Der Einfluss leistungskonformer und leistungsexterner Prüfungsbedingungen auf die Notengebung an deutschen Hochschulen: eine empirische Untersuchung der langfristigen Entwicklung von Examensnoten

Gaens, Thomas

Veröffentlichungsversion / Published Version
Dissertation / phd thesis

DER EINFLUSS LEISTUNGSKONFORMER UND LEISTUNGSEXTERNER PRÜFUNGSBEDINGUNGEN AUF DIE NOTENGEBUNG AN DEUTSCHEN HOCHSCHULEN

Eine empirische Untersuchung der langfristigen Entwicklung von Examensnoten

An der Europa-Universität Flensburg angenommene und mit summa cum laude bewertete Dissertation zur Erlangung des akademischen Grades Dr. phil.

Großer Dank für die Unterstützung des Forschungsprojekts "Hochschulnoten" und dieser Dissertation

- an alle Archivleiter*innen und -mitarbeiter*innen der Hochschulen des samples sowie an alle weiteren Mitarbeiter*innen dort, die die Datenaufnahme ermöglicht haben,
- an meine Kolleginnen Elena Tsarouha, Florence Baillet und Marita McGrory,
- an meine beiden Betreuer Prof. Dr. Volker Müller-Benedict und Prof. Dr. Gerd Grözinger - nicht nur für den fachlichen Austausch!

Shout-outs gehen an:

- meine family im Westen: Joey, Timmy, Svenjamin, Pablo und Schnubbi: Dafür, dass ich auch gefühlte Erdumrundungen von euch entfernt immer auf euch zählen kann,
- meine family in Berlin: Mitja, Patso und Säsch: Dafür, dass ihr seid wie ihr seid,
- Andi, Christian, KLaaRs, Küppi, Matthi, Nina, Nora und Urte: Für eure Freundschaft in all den Jahren und den vielen Spaß,
- Hilke: Für die vielen schönen Zeiten,
- Petra: Für die Reflektion, die Hoffnung, die Sinnlosigkeit, dies, das ;-)
- Christoph, Jonas, Joscha, Laura, Marcus, Mitja, Patrick und Verena: Dafür, dass ihr mich aufgenommen habt, als ich nicht wusste, wohin,
- Mitja: Für so viel und für alles andere!
- Sabina: Dafür, dass mir die Worte fehlen ♥ t. v. & w. s. ♥

worst come to worst - my peoples come first

Inhalt

Tabellen

Abbildungen

1. Einleitung

Die individuellen Lebenschancen innerhalb moderner Gesellschaften sind maßgeblich vom Bildungserfolg des einzelnen Menschen abhängig. In formalisierten Ausbildungssystemen wird anhand standardisierter Bildungszertifikate jede durchlaufene Qualifikationsstufe und die in dieser Stufe erbrachte Leistung dokumentiert. Anhand von Bildungszertifikaten lässt sich beim Eintritt in den Arbeitsmarkt idealerweise die Eignung für verschiedene Positionen und ein bestimmtes Spektrum an Aufgaben feststellen, so dass schon vor der Besetzung von Stellen absehbar ist, ob die Kandidat*innen den zukünftigen Anforderungen gewachsen sein werden. Die konkrete Leistung, die beim Erreichen einer relevanten Qualifikationsstufe, wie etwa der mittleren Reife, der Hochschulreife oder dem Hochschulabschluss erbracht wird, wird in Form von Abschlussnoten festgehalten und ermöglicht eine Differenzierung zwischen den Absolvent*innen derselben Stufe. Auf diese Weise entsteht eine doppelte Rangfolge in den formal nachweisbaren Qualifikationen derjenigen, die das Ausbildungssystem verlassen: Zunächst nach höchstem erreichten Abschluss, dann nach Niveau des Abschlusses, also über die Abschlussnoten und gegebenenfalls über Zusatzleistungen.

Soweit lässt sich die Selektionsfunktion von Bildungszertifikaten in aller Kürze zusammenfassen. Doch werden sie ihrer Aufgabe auch gerecht? Spiegeln die formal abgebildeten Differenzen in Abschlüssen und Noten auch das tatsächliche Leistungsgefälle zwischen Absolvent*innen wider, wie es das meritokratische Prinzip verlangt? Und wie stabil ist diese Vergleichbarkeit im Zeitverlauf? Vor allem für die Leistungsbewertung durch Noten dürften erste Zweifel daran bereits intuitiv aus Erfahrungen mit der Notengebung in der eigenen Bildungskarriere aufkommen. Wer hat nicht schon einmal das Gefühl gehabt, selbst nicht entsprechend der erbrachten Leistung bewertet worden zu sein oder fand die Noten von Kommiliton*innen unangemessen für deren Leistung? Solche Bewertungen sind jedoch stets subjektiv und nicht ausreichend, die Aussagekraft von Noten anzuzweifeln.

Gehaltvoller werden Zweifel durch den Nachweis, dass Noten bereits vor ihrer Nutzung als Differenzierungskriterium beim Eintritt in den Arbeitsmarkt, nämlich als Differenzierungskriterium beim Übergang zwischen den einzelnen Stufen des Bildungssystems nur eingeschränkt ihrem Anspruch gerecht werden. So stellen beispielsweise Abiturnoten den besten Prädiktor für die Erfolgschancen im Studium wieder (Trapmann et al. 2007), allerdings ist auch ihre Vorhersagekraft begrenzt (Müller-Benedict 2010), was nicht weiter verwunderlich ist, lässt sich doch zeigen, dass die gleichen Noten im Abitur an verschiedenen Schulen nicht auch zwangsläufig die gleichen Leistungen abbilden (Neumann et al. 2009).

Besondere Relevanz besitzt die Frage nach der Aussagekraft von Noten in funktional ausdifferenzierten Gesellschaften im Bereich der Hochschulen. Das deutsche Hochschulsystem produziert momen-

tan mehr als 400 000 hochqualifizierte Absolvent*innen pro Jahr[1], die zum Erhalt und zur Weiterentwicklung der hiesigen Wissensgesellschaft beitragen.

Und auch für Hochschulen scheint zu gelten, dass die Aussagekraft von Noten durch beschränkte Vergleichsmöglichkeiten limitiert wird. Diese bisher vor allem aus ungleichheitstheoretischer Perspektive betrachtete Problematik ist nicht erst seit 2003 bekannt, als der Wissenschaftsrat erstmals mit einer Untersuchung der Noten an deutschen Hochschulen der Jahre 1996, 1998 und 2000 an die Öffentlichkeit trat. In dieser stellte das Gremium fest, dass von Fach zu Fach und auch zwischen verschiedenen Abschlussarten deutlich unterschiedliche Notenverteilungen existieren (Wissenschaftsrat 2003). Schon 1987 setzten sich Hitpass und Trosien mit diesem Thema auseinander. Sie konnten ähnliche Unterschiede in der Vergabe von Noten nachweisen (Hitpass/Trosien 1987). In den Jahren 2007 und 2012 bestätigte der Wissenschaftsrat seine Erkenntnisse des Jahres 2003 für den jüngeren Zeitraum der Notengebung weitestgehend (Wissenschaftsrat 2007; 2012) und wies dabei 2007 erstmals dezidiert auf hochschulspezifische Unterschiede innerhalb von Fächern bzw. Studiengängen[2] hin (ebd. 2007).

Müller-Benedict und Tsarouha (2011) systematisierten die bisherigen Ergebnisse zum Thema und hoben in eigenen Analysen der Prüfungsstatistik und anhand der Daten von Hitpass und Trosien die relevanten Differenzierungslinien (Fach-, Abschluss- und Hochschulebene) in der Notengebung hervor. Damit sind nun bereits fast 30 Jahre vergangen, seit mit der Studie von Hitpass und Trosien erstmals Examensnoten an Hochschulen in bundesweitem Umfang fach- (zum Teil studiengang-) und hochschulübergreifend untersucht wurden. Dennoch ist weder die langfristige Entwicklung der Noten, noch deren Vergleich im Querschnitt bisher wirklich aufschlussreich analysiert wurden. Vielmehr sind die verfügbaren Erkenntnisse in zweifacher Hinsicht limitiert. Zum einen konnten auf deskriptiver Ebene zwar erklärungsbedürftige Unterschiede in der Notengebung im Querschnitt aufgezeigt werden, allerdings stehen im Längsschnitt nur unzureichende Informationen zur Verfügung, um zuverlässige Aussagen über den langfristigen Verlauf der Noten zu treffen. Zum anderen fanden sich aufgrund der Beschränkung der Wissenschaftsratsberichte auf deskriptive Darstellungen auf erklärender Ebene lange Zeit lediglich einige Verweise auf die Aussagen von befragten Lehrenden in der Studie von Hitpass und Trosien, die allerdings eher Hypothesencharakter besitzen. Müller-Benedict und Tsarouha begannen 2011 mit ersten weitergehenden Analysen zu potentiellen Einflussfaktoren auf die Notengebung (Eingangseignung; Selektionsneigung von Prüfenden nach Arbeitsmarktlage - genauer in Kapitel 6) damit, diese Forschungslücke zu schließen.

[1] Im Prüfungsjahr 2014 schlossen 432 356 Prüflinge ihr Studium erfolgreich ab (Hochschulrektorenkonferenz 2015).

[2] Um eine beliebige Verwendung dieser Begrifflichkeiten zu vermeiden wird im Folgenden differenziert zwischen a) Fach/Disziplin: Inhaltlich abgrenzbares Teilgebiet der Wissenschaft (z.B. Mathematik) und b) Studiengang: Das Studium eines Fachs mit einem bestimmten Abschluss (z.B. Mathematik Diplom oder Mathematik Lehramt).

Eine erste systematische Erschließung des internationalen wie des Forschungsstandes im deutschsprachigen Raum sowie die empirischen Ergebnisse der Autor*innen implizieren dabei, dass es notwendig ist, die unterschiedlichen Ebenen (Fächer, Abschlüsse, Hochschulen), auf denen sich im Notenniveau Unterschiede in Längs- wie Querschnitt zeigen, auch differenziert zu betrachten.

Im Anschluss an diese bisher umfassendste fach-/studiengang- und hochschulübergreifende systematische Auseinandersetzung mit der Notengebung an deutschen Hochschulen verfolgt die vorliegende Arbeit zwei Ziele. Zum einen wird anhand einer umfassenden deskriptiven Darstellung der seit 1960 vergebenen Abschlussnoten in bis zu 12 Studiengängen an sieben untersuchten Hochschulen ein Überblick über die langfristigen studiengang-, hochschul- und teils auch abschlussspezifischen Benotungsmuster gegeben. Die Examensnoten werden dazu sowohl über verschiedene Zeiträume aggregiert im Querschnitt, als auch im Längsschnitt betrachtet.

Auf diese Weise sind eine quantitative Bestimmung der vermuteten Unterschiede und Gemeinsamkeiten sowie deren Entwicklung möglich. Zentrale Fragen die dadurch beantwortet werden können sind etwa: In welchen Ausmaß unterscheiden sich die durchschnittlichen Notenniveaus zwischen Studiengängen zu gegebenen Zeitpunkten? Wie entwickeln sich die studiengangspezifischen Noten über die Zeit und wie beeinflusst diese Entwicklung die Niveauunterschiede? Gibt es Unterschiede in Niveau und Entwicklung der Noten an einzelnen Hochschulen innerhalb von Studiengängen?

Zum anderen werden potentielle systematische Einflussfaktoren auf die Vergabe von Abschlussnoten näher bestimmt, als bisher geschehen und ihr möglicher Einfluss, soweit dazu geeignete Daten vorliegen, statistisch überprüft. Welche Faktoren weisen Einfluss auf und (wie) unterscheiden sie sich? Um dies zu beantworten, werden mögliche Einwirkungen auf die Notengebung zunächst systematisiert und dahingehend charakterisiert, ob es sich bei ihnen um leistungskonforme oder leistungsexterne Einflüsse handelt, die einmalig oder kontinuierlich Wirkung aufweisen. Es wird eine makrosoziologische Perspektive eingenommen, die den Fokus auf Prüfungsbedingungen legt, die im Aggregat zwischen Fächern und innerhalb dieser zwischen Abschlüssen und Hochschulen in unterschiedlicher Ausgestaltung vorliegen und damit Unterschiede zwischen diesen Analyseebenen bewirken. Dieser vergleichende Blickwinkel wird mit einer Betrachtung im Längsschnitt kombiniert, in der auch mögliche Erklärungen der langfristigen Entwicklung von Notenniveaus identifiziert werden.

Die Arbeit stellt damit Grundlagenforschung dar und weist zu größerem Teil beschreibenden als erklärenden Gehalt auf. Im Anschluss an die Deskription werden einzelne mögliche Ursachen für die aufgezeigten Unterschiede und Entwicklungen auf ihren Einfluss hin überprüft - diese Analysen können jedoch nur den Anfangspunkt einer weitergehenden Erklärung relevanter Mechanismen der Notengebung im Zeitverlauf darstellen, da die Verfügbarkeit geeigneter Daten im Längsschnittformat zum heutigen Zeitpunkt stark begrenzt ist.

Sowohl im deskriptiven als auch im empirischen Teil der Arbeit werden Forschungshypothesen zu Strukturen und Entwicklungen von Examensnoten sowie zu Charakteristika ihrer Beeinflussung überprüft, die aus dem bisherigen Stand der Forschung (Kapitel 2-4) abgeleitet werden.

Um Einflüsse auf den Notengebungsprozess plausibel herleiten zu können, wird dieser zu Beginn der Arbeit einer analytischen Betrachtung unterzogen. Die Notengebung wird als Ergebnis eines Prozesses aus Prüfungs- und Bewertungsphase, der verschiedene inhärente wie externe Anschlussstellen für Verzerrungen der eigentlich abzubildenden Leistung der Geprüften beinhaltet, verstanden und modelliert (Kapitel 2). Aufgrund stabiler Unterschiede in der fachkulturellen Sozialisation der Prüfer*innen ist aus kulturtheoretischer Perspektive eine fachspezifisch stabile Ausgestaltung der Notengebung zu erwarten (Kapitel 3). Die parallel zur fachkulturellen Sozialisation erfolgende Einbindung in ein hochschulspezifisches Setting aus lokalen Prüfungssystemen und lokaler Prüfer*innengemeinschaft bietet allerdings Raum für hochschulspezifische Abweichungen von fachspezifischen Notengebungsmustern. Außerdem lassen sich aus der gesellschaftlichen Funktion der Hochschule als Wissens- und Qualifikationsproduzentin in einer funktional differenzierten Gesellschaft sowie aus ihrer Arbeitsweise als „besondere Organisationen“ (Musselin 2007:63) Annahmen ableiten, die eine Dynamik von Notengebungsmustern im Zeitverlauf erwarten lassen (Kapitel 4).

Im Anschluss an die Formulierung theoretischer Erwartungen an die empirisch vorzufindenden Notengebungsmuster und ein kurzes Zwischenfazit (Kapitel 5) erfolgt eine Auseinandersetzung mit dem bisherigen Stand der Forschung zur Notengebung an Hochschulen. Neben deskriptiven Darstellungen zu Notenniveaus und deren Entwicklungen in Deutschland werden erklärende Ansätze und deren empirische Überprüfungen - aufgrund der marginalen hiesigen Thematisierung vorwiegend aus der internationalen, vor allem aus der US-amerikanischen Literatur – zusammengefasst, kritisch diskutiert und im zuvor erarbeiteten theoretischen Rahmen verortet (Kapitel 6).

Im empirischen Teil der Arbeit wird die Entwicklung der Examensnoten der letzten Jahrzehnte auf Studiengang- und Hochschulebene differenziert dargestellt (Kapitel 8). Dabei steht die quantitative Bestimmung von Unterschieden im Notenniveau zwischen den verschiedenen Analyseeinheiten im aggregierten Querschnitt sowie im Längsschnitt im Vordergrund. Dazu werden Daten aus dem von der Deutschen Forschungsgemeinschaft (DFG) geförderten Forschungsprojekt „Die Notengebung an Hochschulen in Deutschland von den 1960er Jahren bis heute. Trends, Unterschiede, Ursachen.“ Genutzt, welche Zeitreihen der Examensnoten ausgewählter Studiengänge an sieben deutschen Hochschulen enthalten. Detaillierte Informationen zur Erhebung und Aufbereitung der verwendeten Daten finden sich zu Beginn des empirischen Teils, in Kapitel 7.

Die erklärende Analyse orientiert sich an den erklärungsbedürftigen Phänomenen, die sich aus der deskriptiven Betrachtung der Notengebung ergeben. Hier werden die im theoretischen Teil hergeleiteten potentiellen Ursachen für Unterschiede im Notenniveau sowie für dessen langfristige Entwick-

lung auf ihren Erklärungsgehalt für die Notengebung an deutschen Hochschulen hin überprüft, soweit die verfügbaren Daten dies zulassen (Kapitel 8). Abschließend wird noch einmal zusammengefasst, wie sich die studiengang- und standortspezifischen Noten an deutschen Hochschulen in den letzten 50 Jahren entwickelt haben und welche Faktoren zur Erklärung dieser spezifischen Entwicklungen und ihrer Differenzen beitragen können (Kapitel 9).

2. Ein analytisches Modell der Notengebung

2.1 Notengebung als zweistufiger Prozess

Hochschulabschlussprüfungen stellen ein „rechtlich gebundenes und formalisiertes Verfahren zur Feststellung von Fachwissen und fachlicher Leistungsfähigkeit“ (Salzwedel 1996:732) dar. Zur Umsetzung dieses Verfahren müssen die beiden Komponenten Fachwissen und fachliche Leistungsfähigkeit bei jedem Prüfling a) gemessen und b) beurteilt werden. Die Unterteilung der Notengebung in diese zwei Phasen ist deshalb notwendig, weil in ihnen unterschiedliche Prozesse ablaufen: Mittels der Prüfung wird eine Messung durchgeführt, mittels der Bewertung wird die gemessene Prüfungsleistung beurteilt[3] (Rauschenberger 1999; Rheinberg 2002).

Im Fokus jeder im Bildungssystem stattfindenden Prüfung steht dabei die „Vergewisserung des Gelernten“ (Rauschenberger 1999:59), also ob zuvor vermittelte Kenntnisse und/oder Fähigkeiten vom Prüfling internalisiert wurden und wiedergegeben sowie gegebenenfalls angewandt werden können. Diese Überprüfung dient nicht dem Selbstzweck, es wird nicht um der Messung willen gemessen. Mit jeder Prüfung und ihrem Ergebnis ist eine Entscheidung verknüpft. Dabei ist zu berücksichtigen, ob es sich um formative oder summative Prüfungen handelt. Formative Prüfungen begleiten den Lernprozess und kontrollieren den Lernfortschritt. Es steht vor allem eine Rückmeldewirkung an die Geprüften im Vordergrund. Summative Prüfungen sind (bestimmte Lernabschnitte) abschließende Prüfungen, in denen die Signalwirkung der Beurteilung auf alle möglichen Interessent*innen an der Leistung im Fokus steht (Metzger/Nüesch 2004; Yorke 2008). An formative Prüfungen ist die Entscheidung gekoppelt, wie der Lernprozess weiter gestaltet und gefördert werden kann, um das finale Lehrziel zu erreichen. Im Falle summativer Prüfungen muss entschieden werden, ob und in welchem Maße das Lehrziel erreicht wurde. Diese beiden Entscheidungsprozesse repräsentieren die beiden wesentlichen Funktionen von Hochschulprüfungen in modernen[4] Bildungssystemen: Lenkung und Selektion[5] (Metzger/Nüesch 2004). Rein formative Prüfungen wirken über den Lernprozess und damit über die Leistungsfähigkeit des Prüflings auf die Vergabe von Abschlussnoten ein. Die Leistungsfähigkeit der Prüflinge wird im Folgenden als Ergebnis des Zusammenspiels von Eingangseignung und Leistungsentwicklung während des Studiums aufgefasst. Wie genau Leistungsentwicklung im Studium verläuft

[3] Metzger und Nüesch (2004) unterscheiden drei Phasen der Prüfung: Planung, Durchführung und Auswertung. Planung und Durchführung werden hier zusammen als Prüfung an sich verstanden, während die Auswertung als Bewertung des Überprüften als separater Prozess gefasst wird, der auf den eigentlichen Prüfungsprozess folgt. Oft wird auch der Bewertungsprozess noch weiter in Bewertung und Beurteilung unterteilt, diese „Nuancierungen“ (Jachmann 2003:18) werden für den Zweck dieser Arbeit jedoch nicht als relevant erachtet.

[4] Zur historischen Entwicklung von Prüfungen siehe: Prahl 1974.

[5] Es lassen sich auch differenziertere Überlegungen zu den Funktionen von Hochschulprüfungen finden, in denen etwa die berufliche Platzierung, die wissenschaftliche Nachwuchsrekrutierung oder die Herrschafts- und Sozialisationsstabilisierung hervorgehoben werden (etwa Müller/Bayer 2007). Im Grunde stellen aber auch diese Funktionen immer einen Selektionsprozess dar. Eine Ausnahme bildet die Produktion neuer wissenschaftlicher Erkenntnisse in Abschlussprüfungen (vgl. ebd.).

und welche Faktoren in welchem Ausmaß positiven oder negativen Einfluss auf die Leistung haben, ist nicht Thema dieser Arbeit. Entsprechend stehen im Folgenden nur summative Prüfungen und der sich an diese Form der Prüfung anschließende Selektionsvorgang, die Beurteilung, im Fokus.

Die Beurteilung summativer Prüfungen wird in der Regel anhand von Noten (oder Punkten, die sich in Noten umrechnen lassen) vorgenommen. Noten dienen also primär dazu, die überprüfte Leistung sichtbar zu machen[6]. Die Selektion der Prüflinge erfolgt in dem Moment, in dem ihre Leistung in Noten übersetzt wird. Sobald eine Note festgelegt ist, ist damit gleichzeitig festgelegt, a) ob der Prüfling die Prüfung bestanden oder nicht, b) wie gut sein Ergebnis im Vergleich zu anderen Prüflingen ist und c) wie gut sein Ergebnis im Vergleich zu extern gesetzten Maßstäben (etwa Einstellungsvoraussetzungen, Zulassungsbeschränkungen für ein Promotionsstudium etc.) ist.

Wird diese Einordnung des Prüflings durch einen Faktor beeinflusst, der nicht seine Leistung abbildet, muss das Prüfungsergebnis als verzerrt eingestuft werden. In den letzten Jahrzehnten wurde für den deutschsprachigen Raum immer wieder belegt, dass der individuelle Notengebungsprozess in der Schule durch zahlreiche derartige Faktoren beeinflusst wird, die Abbildung von Leistung in Prüfungen entsprechend sensibel vorgenommen und bewertet werden sollte (etwa Ingenkamp 1995 oder Kronig 2007 für zusammenfassende Darstellungen). Noten bilden nicht nur die Leistung selbst ab, sondern beinhalten zusätzlich die Effekte von Messfehlern, individuellen Beurteilungsmaßstäben und Fehlern in der Beurteilung (Abb.1).

Abbildung 1: Analytische Darstellung des individuellen Notengebungsprozesses

2.2 Erste Stufe: Messung

Die Messung, also der Prüfungsprozess, an Hochschulen soll „regelbar, berechenbar und nachprüfbar“ (Salzwedel 1996:732.) sein, um die Vergleichbarkeit der Ergebnisse, das heißt, der Noten, zu gewährleisten. Dabei kommen vor allem testtheoretische Überlegungen zum Tragen, die hervorheben, dass summative Prüfungen valide, reliabel, objektiv, gerecht und ökonomisch gestaltet werden müssen (etwa Linn/Baker/Dunbar 1991; Metzger/Nüesch 2004; Pospeschill 2010; Stieler 2011).

[6] Noten werden zahlreiche weitere Funktionen zugeschrieben, so dienen sie etwa als Ergebnisse formativ wie summativ eingesetzter Prüfungen (z.B. Modulprüfungen im Bachelor/Master System) dazu, Lernfortschritte (quantitativ) zu dokumentieren, Anreize für Leistung zu setzen (Krampen 1984) oder institutionelle Erwartungen an die Lernenden zu kommunizieren (Gamson 1967).
Häufig werden auch Noten und nicht Prüfungen als eigentliches Selektionsinstrument aufgefasst, da Selektion in Notensystemen erst durch die quantifizierende Abbildung von Prüfungsergebnissen realisiert wird (etwa Ratzki 2003).

Genügt die Prüfung nicht den drei zentralen Anforderungen Objektivität, Reliabilität und Validität, kann das Ergebnis einer Prüfung in Form der vergebenen Note bereits durch leistungsunabhängige Faktoren beeinflusst werden, bevor der Prüfling auch nur zur Prüfung angetreten ist.

Objektivität bedeutet in diesem Zusammenhang „die möglichst weitgehende Unabhängigkeit des Messvorgangs von der Subjektivität des Messenden“ (Jäger 2001:207). Reliabilität bezieht sich auf die Zuverlässigkeit der Messung, hinsichtlich der Unabhängigkeit von Rahmenbedingungen des Messvorgangs, etwa von zeitlichen und räumlichen Unterschieden. Validität schließlich bezeichnet die Güte des Messvorgangs in Bezug auf die Abbildung der zu messenden Eigenschaft. Der Grad der Validität kann übersetzt werden mit dem Grad, in dem es der Messung gelingt, tatsächlich nur die interessierende Eigenschaft zu messen (ebd.; zur weiteren Differenzierung innerhalb der einzelnen Anforderungen im Prüfungsprozess siehe Lienert/Raatz 1998). Hinzuweisen ist auf den Umstand, dass die beschriebenen Anforderungen ein Idealbild zeichnen, das in der Praxis nicht in optimaler Ausführung erreicht werden kann – ein Umstand, der gelegentlich schon als ausreichend betrachtet wird, um die Vergleichbarkeit von Noten über einzelne Organisationseinheiten hinaus in Frage zu stellen (etwa Webler 2010).

Doch auch wenn die Gütekriterien der Testkonstruktion nicht in vollem Umfang umsetzbar sind, bedeutet dies nicht, dass sie als Wunschdenken abgetan werden können. Prüfungen lassen sich in Bezug auf diese Anforderungen graduell unterscheiden - weist die Testkonstruktion bereits vergleichsweise erhebliche Mängel in Objektivität, Reliabilität und Validität auf, ist die Leistungsabbildung von vorneherein in fahrlässiger Weise beeinträchtigt.

Verschiedene Studien konnten bereits belegen, dass in allen Bereichen des primären und sekundären Bildungssystems testtheoretische Einschränkungen, die zu einer verzerrten Abbildung von Leistung führen, eher den Normalfall, als eine Ausnahme darstellen (siehe die Überblicke in Jäger 2001; Lissmann 1997). Und auch im von der Forschung vergleichsweise vernachlässigten Bereich der tertiären Bildung existieren Befunde, die darauf hinweisen, dass die gewünschte stabile und genaue Leistungsmessung in Prüfungen nur eingeschränkt gelingt (etwa Novy et al. 1996; Wass et al. 2003). Im Hinblick auf die hier im Fokus stehende Frage nach der Vergleichbarkeit von Noten - im Querschnitt wie im Längsschnitt - gilt es deshalb zu klären, inwiefern sie durch testtheoretische Einschränkungen beeinflusst werden kann. Dass Prüfungen die abgefragte Leistung nicht zu 100% abbilden können, kann sich im Aggregat in mehrfacher Hinsicht auswirken. Zunächst einmal muss die Entwicklung der Noten im Zeitverlauf in Abhängigkeit zur Entwicklung der Testkonstruktion betrachtet werden. Durch die stete Evaluation von Prüfungen durch die testtheoretische Forschung kann davon ausgegangen werden, dass die Testinstrumente sich weiterentwickelt haben. Es ist wahrscheinlich, dass Prüfungen im Zeitverlauf in zunehmendem Maße die zu messende Leistung abbilden. Unter der Kontrolle von Leistung wäre in diesem Fall eine im Zeitverlauf abnehmende Streuung der Noten zu erwarten. Bes-

sere Noten[7] könnten dann möglicherweise darauf zurückgeführt werden, dass die schon immer gute Leistung der Studierenden heute einfach besser erfasst werden kann.

In Bezug auf die Vergleichbarkeit von Noten im Querschnitt sind die unterschiedlich hohen Gütegrade verschiedener Prüfungsformen von Relevanz: Standardisierte Tests weisen eine höhere Objektivität und Reliabilität auf als schriftliche (aufsatzartige) Prüfungen, mündliche Prüfungen erzielen hier die niedrigsten Werte (siehe den Überblick in Birkel 1978 - auch ausführlich zur Güte mündlicher Prüfungen - für schriftliche Arbeiten die Untersuchungen in Ingenkamp 1995).

Möglicherweise können Unterschiede im Notenniveau zwischen Fächern, Abschlussarten oder Hochschulen also zu einem Teil darauf zurückgeführt werden, dass die dort jeweils durchgeführten Prüfungen sich zu unterschiedlichen Teilen aus mündlichen und schriftlichen Prüfungen zusammensetzen, die mündliche und schriftliche Leistung unterschiedlich gewichtet in die Gesamtnote eingeht. Mündliche Prüfungen könnten zudem eine größere Streuung aufweisen als schriftliche, da die Prüfungssituation als mündliche Interaktion zwangsläufig einen geringeren Standardisierungsgrad aufweist als schriftliche Arbeiten. Eine größere Bandbreite in den Interaktionsspielräumen könnte zu einer größeren Bandbreite in der Notenvergabe führen. Neben der Frage, wie genau die abgefragte Leistung in Prüfungen gemessen wird, muss außerdem berücksichtigt werden, was für eine Art von Leistung eigentlich abgefragt wird. Faktenwissen etwa kann in mündlichen wie auch in schriftlichen Prüfungen eindeutiger und schneller als falsch oder richtig eingestuft und entsprechend bewertet werden als Interpretationen. Während Interpretationen in schriftlich vorliegenden Arbeiten allerdings unter ausführlicher und mehrmaliger Betrachtung in Ruhe einem Urteil unterzogen werden können, verlangt die mündliche Prüfung eine wesentlich zeitnähere Bewertung.

Schließlich lösen mündliche Prüfungen als Stresssituation mit ihrer punktuellen Zuspitzung der Leistungsabfrage eher Prüfungsangst aus als schriftliche Hausarbeiten und lassen zudem keine externen Hilfen (wie etwa korrekturlesende Dritte) zu. Dies spricht dafür, dass mündliche Prüfungen die tatsächlichen Fähigkeiten eines Prüflings konservativer messen als Hausarbeiten. Hausarbeiten und schriftliche Prüfungen in Klausurform sollten Leistung somit genauer abbilden als mündliche Prüfungen. Die Noten in mündlichen Prüfungen sollten unter Kontrolle von Leistung (bei tatsächlich gleicher Leistung) stärker streuen, da mehr Interpretationsspielraum für die Prüfenden besteht. In Hausarbeiten kann die eigene Leistung am ehesten ausgereizt werden, da keine örtlichen Beschränkungen (Verzicht auf Hilfsmittel und Aufzeichnungen, externe Dokumente und Quellen) bestehen und externe Korrekturen nicht verhindert werden können. In diesem Format sind eher Verzerrungen der Leistung zum Besseren zu erwarten.

[7] *Bessere* Noten sind im deutschen Bildungssystem gleichbedeutend mit *sinkenden* Noten, da die deutsche Notenskala in der Regel von 1 (‚sehr gut') bis 6 (‚ungenügend'), gelegentlich auch nur bis 5 (‚mangelhaft') reicht, wobei die 4 (‚ausreichend') das niedrigste Resultat darstellt, bei dem eine Prüfung als bestanden gilt.

2.3 Zweite Stufe: Bewertung

Ist die Prüfung vom Prüfling beendet, beginnt die zweite Phase der Notengebung, in der es gilt, das Messergebnis, also den unternommenen Versuch zum Nachweis von Fachwissen und fachlicher Leistungsfähigkeit in der Abschlussprüfung zu beurteilen. Um selektiv wirksam zu sein, müssen die Prüfenden zum einen entscheiden, ob ein Mindestmaß an Wissen und Leistungsfähigkeit erfüllt ist, ob der Abschluss also erteilt wird oder nicht, zum anderen müssen sie entscheiden, in welchem Maße Wissen und Leistungsfähigkeit vorliegen. Diese Selektion erfolgt mittels Notenvergabe. Jedoch ergeben sich die Anhaltspunkte und Kriterien für die Benotung nicht aus der Prüfung an sich, sondern aus der Setzung von Maßstäben. Erst durch die Setzung von Maßstäben können Prüfungsleistungen in Noten (oder zunächst in Punkte) übersetzt werden. Den Prüfenden stehen dazu unterschiedliche Bewertungsmaßstäbe zur Verfügung, sogenannte Bezugsnormen.

Das Konzept der Bezugsnormen geht auf Heckhausen (1974) zurück und geht von der Notwendigkeit aus, beobachtbares Verhalten einordnen zu müssen, um es als Leistung beurteilen zu können. Zur Beurteilung von Verhalten stehen drei Bezugsnormen zur Verfügung: Die individuelle, die soziale und die sachliche Bezugsnorm[8]. Die individuelle Bezugsnorm stellt einen intra-individuellen Vergleich dar und ermöglicht den Vergleich einer bestimmten Leistung mit einer Leistung derselben Person zu einem früheren Zeitpunkt. Mithilfe der sozialen Bezugsnorm wird die Leistung einer Person mit der Leistung einer anderen, also inter-individuell, verglichen. Die sachliche Bezugsnorm schließlich ermöglicht es, Leistung mit festgelegten, zuvor zu definierenden Leistungskriterien zu vergleichen (Heckhausen 1974). Lehrer*innen bewerten die Leistungen ihrer Schüler*innen demnach entsprechend individueller Lernfortschritte (individuelle Bezugsnorm), der Leistungen der anderen Schüler*innen (soziale Bezugsnorm) sowie der im Lehrplan festgelegten Lehrziele (sachliche Bezugsnorm, auch: curriculare Bezugsnorm). Die Bezugsnormen werden dabei kombiniert angewendet. Rheinberg konnte jedoch feststellen, dass Lehrer*innen in der Regel eine bestimmte Bezugsnormorientierung aufweisen, einige also bei ihren Bewertungen neben der sachlichen Bezugsnorm unabhängig vom jeweiligen Bewertungskontext grundsätzlich eher die soziale, andere grundsätzlich eher die individuelle Bezugsnorm anwenden (Martinek 2007; Rheinberg 1980; 2002).

Welche Bezugsnorm bei der Notenvergabe angewendet wird, hat dabei deutlichen Einfluss auf die Beurteilung. Jede der drei Normen ist mit Einschränkungen in Hinblick auf die selektive Aufgabe der Notenvergabe belastet. Gänzlich ungeeignet für die Vergabe von Abschlussnoten ist die individuelle Bezugsnorm. Ihre Anwendung bildet lediglich den Lernfortschritt eines Prüflings ab, also ob ein Prüfling sein Fachwissen und seine fachliche Leistungsfähigkeit steigern konnte oder nicht und ist damit

[8] Heckhausen führt zudem als vierte Bezugsnorm die fremdgesetzte Norm an, die sich von der sachlichen Bezugsnorm allerdings lediglich dahingehend unterscheidet, dass die als Maßstab dienenden Kriterien der Leistungserreichung nicht vom Beurteilenden selbst, sondern von „dafür zuständigen Instanzen“ (Heckhausen 1974:51) bestimmt werden.

vor allem als Motivationsinstrument geeignet. Sie bietet aber weder die Möglichkeit festzustellen, ob das geforderte Mindestmaß an Kompetenzen vorliegt, noch, in welchem Maße dies der Fall ist (Metzger/Nüesch 2004; Rheinberg 2002). Die sachliche Bezugsnorm scheint intuitiv einer objektiven Leistungsmessung angemessen, misst sie doch die erbrachte Leistung an den erforderten Lehrzielen. Zuvor klar definierte Vorgaben, welches Maß an fachlichem Wissen und Leistungsfähigkeit vorliegen muss, um die Prüfung zu bestehen, erlauben für jeden Prüfling prinzipiell eine zweifelsfreie Bewertung, ob diese Vorgaben erfüllt wurden oder nicht. Die sachliche Bezugsnorm erfüllt damit den Anspruch des Prüfungsverfahrens zu beurteilen, ob geforderte Kompetenzen vorliegen oder nicht. Die alleinige Anwendung dieser Norm bringt allerdings ein Problem bei der Notenvergabe mit sich: Es kann mittels ihrer Hilfe nur entschieden werden, ob die eingangs gesetzten Vorgaben erfüllt wurden oder nicht, das heißt, ob die Prüfung bestanden wurde oder nicht – was ihre Anwendung bei der Vorgabe eines zu erreichenden Kompetenzniveaus allerdings nicht zu leisten vermag, ist eine Einschätzung darüber abzugeben, wie gut diese Vorgaben erfüllt wurden oder nicht erfüllt wurden. Der alleinige Bezug auf sachliche, curriculare Kriterien, die einen bestimmten Mindeststand an Kompetenzen abbilden sollen, selektiert die Prüflinge also lediglich in zwei Gruppen und bildet deren jeweiligen Kompetenzstand ab, erlaubt aber noch keine Anordnung der Prüfungsleistungen in eine Rangfolge entsprechend einer Notenskala. Prinzipiell besteht zwar die Möglichkeit, die zu erreichenden sachlichen Kriterien vor der Prüfung in verschiedene Stufen zu gliedern und jede dieser Stufen mit einer Note zu verknüpfen, so dass für ein bestimmtes Mehr an fachlichen Kompetenzen eine bessere Note erteilt wird. Ein solches Verfahren birgt allerdings die Problematik, dass zur Gewährleistung der Vergleichbarkeit der Noten jede Prüfung den gleichen Schwierigkeitsgrad mit der exakt gleichen Differenz zwischen Kompetenzunterschieden für die einzelnen Notenstufen aufweisen müsste (vgl. Heckhausen 1974; Martinek 2007; Metzger/Nüesch 2004; Rheinberg 2002).

Dies scheint schon durch den Umstand unwahrscheinlich, dass in tertiären Bildungseinrichtungen nicht einmal eine einheitliche Vorstellung davon existiert, wie Bezugskriterien bzw. Bezugsstandards formal zu definieren sind, ganz abgesehen davon, dass weder vergleichbare Verfahren noch eine inhaltliche Übereinstimmung von Standards gewährleistet sind (Sadler 2005). Die Vergleichbarkeit von Prüfungsleistungen wird in der Praxis stattdessen in der Regel durch die Anwendung der sozialen Bezugsnorm erreicht. Erst die Gesamtheit der Prüfungsergebnisse zeigt, welche Leistung im konkreten Fall der Leistungsmessung durchschnittlich überhaupt erreicht werden konnte. Durch den Vergleich der einzelnen Prüfungsleistung mit den Leistungen anderer Prüflinge kann dann entschieden werden, wie gut Einzelne die an sie gestellten Aufgaben vergleichsweise erfüllt haben. Nutzen Lehrende vorwiegend die soziale Bezugsnorm wird dadurch eine Normalverteilung der vergebenen Noten begünstigt (Metzger/Nüesch 2004). Eine alleinige Orientierung am Gruppendurchschnitt führt deshalb zum sogenannten Referenzgruppeneffekt: Prüflinge einer Prüfungsgruppe mit einer gegebe-

nen Durchschnittsleistung erhalten bei der Bewertung ihrer Leistung bessere (schlechtere) Noten als Prüflinge mit ähnlicher Leistung in durchschnittlich stärkeren (schwächeren) Prüfungsgruppen (Davis 1966; Ingenkamp 1970; Köckeis-Stangl 1972; Südkamp/Möller 2009; Trautwein/Baeriswyl 2007).
Die Benotung entlang einer normalverteilten Kurve begrenzt die Zahl der Bestnoten, während bei einer vorwiegenden Nutzung der sachlichen Bezugsnorm prinzipiell alle Prüflinge mit der Bestnote bestehen können. Eine zunehmende Nutzung der sachlichen Bezugsnorm könnte also zur Verbesserung von Notenniveaus beitragen (Olsen 1997; Rojstaczer/Healy 2012; Yorke 2011). Zudem mindert die Nutzung der sozialen Bezugsnorm die Vergleichbarkeit von Noten durch den Referenzgruppeneffekt, durch den das mittlere Leistungsniveau von schwachen Prüfungsgruppen überschätzt, das von starken Prüfungsgruppen unterschätzt wird. Die Wahl der Bezugsnorm bzw. unterschiedliche Bezugsnormenorientierungen könnten demnach Einfluss auf die Notengebung besitzen.

Unabhängig davon, welche Bezugsnorm die Lehrenden verwenden, wird die Beurteilung von Prüfungsleistungen von einigen gut dokumentierten Verzerrungsmechanismen begleitet[9]. Einige davon entfalten ihre Wirkung erst ab einem bestimmten Mindestmaß an Kontakt zwischen Prüfling und Prüfer*in bzw. ab einem Mindestmaß an Kenntnissen über bestimmte Eigenschaften des Prüflings. Dazu gehört die Gruppe der Wahrnehmungs- oder auch Inferenzfehler, zu denen der logische Fehler sowie die Phänomene der selektiven Wahrnehmung und der Stereotypisierung zählen (Lissmann 1997; Sacher 1996). Alle drei beschreiben im Grunde den Einfluss jeweils bewusst oder unterbewusst wirkender Annahmen von Prüfer*innen über den Zusammenhang bestimmter Eigenschaften von Prüflingen und deren Leistung, so dass die Kompositionsmerkmale selbst gar nicht im Prüfungsergebnis wirksam werden müssen, um dieses zu beeinflussen (Frauen müssen etwa gar nicht schlechter in Mathematik sein, um schlechtere Noten zu erhalten, wenn der (im Zweifelsfall männliche) Dozent davon überzeugt ist, dass sie im Durchschnitt nun einmal weniger mathematische Fertigkeiten besitzen). Diese Annahmen und die mit ihnen verbundene Urteilstendenz (positiv wie negativ) benötigen im Falle der unterbewussten Wirkung nur einen sehr geringen Bestätigungsgrad in der sozialen Wirklichkeit, damit sie bei der Benotung in Kraft treten. Umgekehrt führen auch starke Evidenzen gegen den implizit angenommenen Zusammenhang nur selten zu dessen Revidierung und damit zu einer

[9] Nicht relevant für die hier im Mittelpunkt stehende Benotung von Hochschulprüfungen sind in der Schulforschung dokumentierte Verzerrungsmechanismen, die nur indirekt über die Leistung der Schüler*innen die Benotung beeinflussen. Ein Beispiel ist die Wirkung des Erwartungseffekts. Dieser beschreibt eine Verknüpfung von Prüflingsmerkmalen mit Leistungserwartungen durch Lehrende. Diese Erwartungen verzerren aber nicht direkt das Urteil von Lehrenden in der Benotung, sondern deren Verhalten gegenüber den Prüflingen, welche (aufgrund positiver Erwartungen) positiv auf die Lehrperson reagieren, wenn sie positiv behandelt werden, während im Falle negativer Erwartungen die darauf folgende negative Behandlung auch entsprechend ablehnend beantwortet wird. Ein solcher Effekt, beeinflusst mittelbar zwar die Benotung, der direkte Zusammenhang zwischen in der Prüfungssituation gemessener Leistung und erteilter Note wird dadurch aber nicht berührt.

Korrektur der vorurteilsbelasteten Benotung. Wird das Urteil durch ein besonders auffälligen Merkmales des Prüflings beeinflusst, wird auch vom Halo-Effekt gesprochen (ebd.). Diese im Bereich der Schulforschung dokumentierten Phänomene sind im Bereich der Hochschulprüfungen vor allem in kleineren Studiengängen zu vermuten, in denen geringe Kursgrößen es ermöglichen, dass die Lehrenden ihre Studierenden öfter beobachten können und tatsächlich in gewissem Maße kennen. Auch in der relativ anonymen Bewertungssituation, die in Massenfächern herrscht, muss allerdings in Betracht gezogen werden, dass der Name des Prüflings Rückschlüsse auf dessen Geschlecht und Ethnizität (und in begrenztem Maße auch auf dessen Schichtzugehörigkeit) zulässt - Eigenschaften, die möglicherweise mit unbewussten Stereotypen oder bewussten Vorurteilen belastet sind. Entsprechend könnte die Zusammensetzung der Studierenden in Bezug auf diese Merkmale auch unabhängig von tatsächlichen Leistungsunterschieden zwischen den Geschlechtern oder Ethnien zu systematischen Unterschieden in der Notengebung führen. Zu prüfen wäre auch, ob ein solcher Effekt, ließe er sich nachweisen, in Fächern mit geringen Studierendenzahlen und guten Betreuungsrelationen stärker ist, als in Fächern mit hohen Studierendenzahlen und schlechten Betreuungsrelationen. Auch fachspezifische Ausprägungen, etwa auf klassischen Geschlechtervorstellungen beruhend, sind denkbare Folgen solcher Wahrnehmungsfehler im Bewertungsprozess, die ein leistungsexternes Äquivalent zum leistungskonformen Einfluss kompositioneller Faktoren darstellen (vgl. Faber/Billmann-Mahecha 2010).

Neben Merkmalen des Prüflings können auch situative Faktoren der Prüfung die Wahrnehmung der Prüfungsleistung und damit deren Beurteilung verzerren. So sind etwa der sogenannte Primacy-Effekt und sein Äquivalent, der Recency-Effekt, dafür verantwortlich, dass sich das Gedächtnis von Prüfer*innen vor allem an den ersten bzw. den letzten Eindruck einer Prüfungsleistung erinnert (Lissmann 1997). Unterschiede im Prüfungsaufbau, etwa ob die einzelnen Prüfungsteile im Laufe der Prüfung zunehmend leichter oder schwerer werden oder ob die Prüfung einen konstanten Schwierigkeitsgrad aufweist, könnten aufgrund dieser Effekte einen systematischen Einfluss auf die Notengebung ausüben. Reihenfolgeeffekte sind jedoch nicht nur bei der Bewertung einzelner Prüflinge zu finden, sie lassen sich auch bei der Beurteilung ganzer Prüfungskohorten nachweisen. So werden Prüflinge, die am Ende eines Prüfungsdurchlaufs geprüft werden oder deren schriftliche Arbeiten als letzte korrigiert werden, nach einer häufig vorgetragenen These stärker unter dem Eindruck der zuvor abgenommenen Leistungen beurteilt, als dies bei den ersten Prüflingen möglich ist (Birkel (1978) berichtet sowohl von dem Befund, nach dem die Noten zunehmend besser werden, als auch von dem Befund schlechter werdender Noten). Auch die Beurteilung ganzer Prüfungsjahrgänge könnte auf diese Weise beeinflusst werden, steigen die Vergleichsmöglichkeiten der Prüfenden doch mit jedem weiteren Prüfungsdurchgang an. In diesem Hinblick ist auch die Erfahrung der Prüfenden eventuell von Bedeutung für die Notenvergabe (vgl. Jäger 2001). Unterschiede zwischen den Hoch-

schulen hinsichtlich der Prüfungserfahrung der dort Lehrenden oder zwischen Studiengängen hinsichtlich der Prüfungsbelastung könnten auf diese Weise systematische Verzerrungen im Notenbild begünstigen.

Betz (1974) berichtet hingegen von einem gegenteiligen Effekt. In einer empirischen Untersuchung von 60 mündlichen Diplom-Vorprüfungen in Psychologie zeigt sich, dass die vergebenen Noten bei allen untersuchten Prüfer*innen unabhängig von der Reihenfolge der Kandidat*innen periodische Schwankungen aufweisen. Diese Schwankungen nehmen mit der Länge des Prüfungsblocks zu, das heißt, die Schwankungsperioden werden mit zunehmender Anzahl Prüflinge kürzer. Betz schlussfolgert, dass sich bei den Prüfenden mit zunehmender Dauer der Prüfungsblöcke eine „Sättigung" (ebd.: 10) einstellt, durch die die Zahl der jeweils zum Vergleich herangezogenen Leistungen zunehmend abnimmt, bis schließlich nur noch die unmittelbar zuvor geprüfte Leistung als Referenz genutzt wird.

Ein stärker bewusst ablaufender Verzerrungsfaktor ist der von Sacher als „Wissen-um-die Folgen-Fehler" (1996:179) bezeichnete Einfluss einer Prüfungssituation, in der Prüfende abschätzen können, welche Bedeutung die von ihnen gegebene Note auf die Zukunft des Prüflings hat. Während im Schulbereich vor allem Fragen des Bildungsübergangs oder der Versetzung im Mittelpunkt stehen, fällt bei einer Übertragung auf den Hochschulbereich vor allem eine beeinflusste Selektionsneigung aufgrund einer bestimmten Arbeitsmarktlage in diese Kategorie (vgl. Müller-Benedict/Tsarouha 2011). Deshalb sollte diese Verzerrungsquelle nicht mehr dem eigentlichen Bewertungsprozess zugeschrieben werden.

In der Literatur ebenfalls aufgeführt werden Merkmale von Prüfenden, die die Vergleichbarkeit der Noten zwischen einzelnen Lehrenden einschränken. Zu ihnen zählen etwa unterschiedliche Attribuierungsmuster von Prüfer*innen. Unterschiede in der Benotung zwischen einzelnen Lehrenden können nach dieser Annahme zum Teil dadurch erklärt werden, dass einige von ihnen die Prüfungsergebnisse den Leistungen der Prüflinge und deren Lernbemühungen zuschreiben, andere die Ergebnisse primär als Ergebnis ihrer Lehrleistung verstehen. Letztere würden ihre Prüfungen vermutlich tendenziell besser bewerten (Lissmann 1997). Um Unterschiede zwischen Fächern, Abschlüssen und/oder Hochschulen - oder auch im Zeitverlauf - erklären zu können, müssten sich die Attribuierungsmuster der Lehrenden allerdings systematisch unterscheiden bzw. entwickeln. Auch hinsichtlich individueller Beurteilungstendenzen müssten sich solche Muster begründen und nachweisen lassen, sollen sie zur Erklärung der Notengebung beitragen. Dies betrifft nachgewiesene Beurteilungsmuster wie Strenge- oder Mildefehler, die eine besonders starke bzw. schwache Selektionsneigung einzelner Prüfer*innen beschreiben oder die Tendenz zur Mitte bzw. zu Extremurteilen, bezüglich der Ausschöpfung des Notenspektrums (Jäger 2001). Fachkulturelle Unterschiede im Bewertungsstil, die einen systematischen Einfluss dieser individuell nachgewiesenen Verzerrungsfaktoren nach sich ziehen sind denkbar,

müssten aber zunächst im Vergleich theoretisch hergeleitet werden. Im Durchschnitt kaum relevant dürften der Kontrast- und der Projektionseffekt sein. Ersterer beschreibt die Tendenz, vor allem Eigenschaften der Prüflinge wahrzunehmen und in der Benotung stärker zu gewichten, die bei den Prüfenden selbst positiv ausgeprägt sind, Letzterer die gegenteilige Tendenz, vor allem selbst vorhandene Schwächen bei den Prüflingen wahrzunehmen (Ziegenspeck 1999).
Zuletzt sind individuelle Merkmale der Prüfungsleistung selbst, für die in der Schulforschung ein Zusammenhang mit der Beurteilungstendenz festgestellt werden konnten, als Verzerrungsfaktoren dokumentiert. In diese Kategorie fallen etwa Vergleichseinordnungen, die berücksichtigen, wie ordentlich die Schrift der Prüflinge ist oder wie lang schriftliche Arbeiten im Vergleich zu Mitprüflingen sind (Jäger 2001), welche sich im Durchschnitt zwischen verschiedenen Analyseeinheiten kaum überzufällig unterscheiden dürften und daher bei der Analyse der Notengebung auf Aggregatebene unberücksichtigt bleiben können.

Die bisherigen Erkenntnisse der Bildungsforschung zur Notengebung, vor allem in Schulen, lassen darauf schließen, dass die Messung und Bewertung von Leistungen selbst bereits durch inhärente Verzerrungsquellen systematische Unterschiede in der Benotung unterschiedlich abzugrenzender Prüflingsgruppen begünstigt. Die Notengebung, verstanden als zweistufiger Prozess, ist demnach nicht nur, wie es die Schulforschung zeigt, grundsätzlich fehleranfällig, was die Benotung individueller Leistung betrifft. Auch auf Aggregatebene muss die Möglichkeit in Betracht gezogen werden, dass Leistung zwischen verschiedenen Organisationseinheiten unterschiedlich genau abgebildet und zum Besseren oder Schlechteren hin verzerrt wird. Im Vergleich zwischen Fächern, Abschlussarten und/oder Hochschulen könnten dafür verantwortlich sein:

- Die unterschiedliche Anwendung und Gewichtung verschiedener Prüfungsformate (Leistung wird in standardisierten Klausuren am genauesten erfasst, während in mündlichen Prüfungen eine stärkere Streuung der Noten um die leistungsäquivalente Note zu erwarten ist; in Hausarbeiten sind die besten Noten zu erwarten)
- Die Zusammensetzung der Studierenden nach stereotypisierenden Merkmalen (z.B. Geschlecht, Ethnizität, soziale Herkunft), wobei ein derartiger Einfluss durch die Anzahl der Prüflinge moderiert werden könnte und in seinem Ausmaß von den Einstellungen der Prüfenden abhängt.
- Die durchschnittliche Prüfungserfahrung der Prüfer*innen (mit der Erfahrung zu- vs. abnehmender Vergleich mit zurückliegenden Prüfungskohorten) und die Höhe der Prüfungsbelastung
- Die (un)bewusste Einschätzung der Auswirkungen von Noten, welche sich aus (der Wahrnehmung von) spezifischen Zukunftschancen ergibt

Im Zeitverlauf könnten folgende Entwicklungen als dem Notengebungsprozess inhärente Auswirkungen auftreten:

- Eine zunehmend genauere Abbildung der tatsächlichen Leistung (durch zunehmend verbesserte Testinstrumente)

- Eine erhöhter Anteil guter Noten (durch die Umstellung von der sozialen auf die sachliche Bezugsnorm)
- Ein abnehmender Einfluss eines möglichen Effekts der sozialen Zusammensetzung von Studierenden (durch steigende Studierendenzahlen und damit zunehmende Anonymität)

Trotz der zu erwartenden Differenzen im Notenniveau beim Vergleich von nach verschiedenen Kriterien (z.B. Fach, Abschluss, Hochschule) zusammengefassten Gruppen von Prüflingen aufgrund leistungsexterner Einflüsse muss an dieser Stelle erwähnt werden, dass davon auszugehen ist, dass ein beachtlicher Teil der Note auf die tatsächliche Leistungsfähigkeit der Studierenden zurückzuführen ist. Diese Annahme ergibt sich aus dem mehrfach abgesicherten Befund, dass die Eingangseignung von Studierenden, hierzulande in der Regel operationalisiert als Abiturnote, den besten einzelnen Prädiktor für die Examensnote darstellt (Trapmann et al. 2007; zusammenfassend: Köller 2013). Bei der Erklärung von unterschiedlichen Notenniveaus muss deshalb immer auch berücksichtigt werden, dass Unterschiede sich möglicherweise einfach durch unterschiedlich gute Leistungen der Studierenden ergeben. Daraus folgt eine grundlegende Forschungshypothese für die Analyse der Notengebung an deutschen Hochschulen:

FH1: Mögliche Unterschiede im Notenniveau zwischen nach Fächern, Abschlüssen und/oder Hochschulen abgrenzbaren Gruppen von Prüflingen sind sowohl auf leistungskonforme, als auch auf leistungsexterne Ursachen zurückzuführen.

3. Fachspezifische Bedingungen der Notengebung

3.1 Notengebung als fachlich eingebettetes soziales Handeln

Wie die einführende Beschreibung des Notengebungsprozesses zeigt, liegen auf dem Weg von der Leistung, die durch eine Note abgebildet werden soll, zur letztendlich vergebenen Zensur einige Stolpersteine, die zu einer verzerrten Messung und/oder Beurteilung der interessierenden Kompetenz(en) führen können. Wie bereits angedeutet, wirken diese Einflüsse aber nicht beliebig auf die Notenvergabe. Wird die individuelle Ebene des Selektionsprozesses, auf der der einzelne Prüfling im Mittelpunkt steht, verlassen und stattdessen der größere Kontext betrachtet, in den er durch die Beurteilung seiner Leistung eingeordnet wird, liegt der Fokus auf der Aggregatebene. Hier wird die Leistung des Individuums und die daraus übersetzte Note zu einem Bestandteil des Notenniveaus, das sich für verschiedene Analyseebenen, denen der Prüfling zugeordnet werden kann, und bestimmte Zeitpunkte oder- räume bestimmten lässt: Das Notenniveau des Faches, in dem er studiert, der Abschlussart in diesem Fach, auf die er studiert, der Hochschule an der er dieses Fach auf diesen Abschluss studiert usw.

Erst *systematische* Unterschiede im Prüfungs- und Bewertungsprozess zwischen Organisationseinheiten oder im Zeitverlauf lassen Verzerrungen für die Analyse von Noten auf der Makroebene interessant werden, da hier eine Systematik entsteht, die sich in den Durchschnittsnoten wiederfinden lässt. Unterscheiden sich die betrachteten Einheiten oder Zeiträume nicht systematisch hinsichtlich der Verzerrungsfaktoren, gleichen sie sich im Aggregat weitestgehend aus und weisen keinen Effekt auf die durchschnittliche Notengebung auf. Dann wären unterschiedliche Notenmuster in Quer- und Längsschnitt lediglich auf leistungskonforme Ursachen zurückzuführen.

Es scheint aber naheliegend, dass systematische Unterschiede auch auf leistungsunabhängigen Faktoren beruhen. Müller-Benedict weist mehrfach auf den möglichen Einfluss von Fachkulturen auf die Notengebung hin: So umfassen Fachkulturen als „Spektrum von in der Fachdisziplin geteilten wissenschaftlichen Anschauungen, Vorgehensweisen und Normen (…) u. a. auch ein fachspezifisches Notenspektrum“ (Müller-Benedict/Tsarouha 2011). Unterschiedliche Notenniveaus in verschiedenen Fächern müssen demnach auch „als Auswirkung der Fachkulturen (…), die auf der unterschiedlichen gesellschaftlichen Funktionalität, den unterschiedlichen historischen Entwicklungspfaden und (…) späteren Eingriffen in die Benotung über Prüfungsordnungen (…) beruhen“ (Müller-Benedict et al. 2008:39) verstanden werden. Die Stabilität fachspezifischer Notenniveaus erklären Müller-Benedict und Tsarouha anhand nach außen relativ geschlossener Fachkommunikation, die Etablierung unterschiedlicher Hochschulstandards innerhalb eines Faches unter anderem mit lokal unterschiedlichen Arbeitsbedingungen der Lehrenden und fakultätsinternen Anpassungsprozessen (dieses Argument findet sich auch bei Agnew (1995)).

Dass sich auch Verzerrungsfaktoren systematisch zwischen akademischen Organisationseinheiten unterscheiden, kann dadurch erklärt werden, dass die Benotung auf lokaler Ebene durch strukturelle Zwänge gerahmt wird. Diese Zwänge manifestieren sich im jeweils vorherrschenden *Prüfungssystem*, dessen Ausgestaltung sich aus unterschiedlichen Bedingungsfaktoren zweier Kategorien und deren jeweiligen Ausprägungen ergibt:

> „Prüfungssysteme werden durch manifeste Normen (Rahmenregelungen, Prüfungsordnungen, Studienordnungen, Studienpläne, Durchführungsbestimmungen) und latente Normen („geheimer Lehrplan", vorherrschende Sozialisationsmuster, örtliche oder fachliche Besonderheiten) gesteuert. Daneben manifestieren sich darin auch Momente wie die übliche Arbeitsform der jeweiligen Disziplin, der Konsens oder Dissens über Inhalte, Methoden und Anwendung der Disziplin oder die quantitative Relation zwischen Lehrenden und Lernenden." (Prahl 1983:441)

Was Prahl hier beschreibt, ist das jeweils konkrete Zusammenspiel von Prüfungsbedingungen, das für alle Prüfungen innerhalb einer bestimmten Organisationeinheit einen einheitlichen Rahmen schaffen soll. Das heißt nichts anderes, als dass sich der eingangs modellierte Notengebungsprozess für jede akademische Organisationeinheit vom einzelnen Studiengang über die Hochschule bis zu einem kompletten Fach als das Produkt bestimmter, für diese Einheit spezifischer Ausprägungen seiner einzelnen Komponenten fassen lässt. Die historische Entwicklung der Bedingungsfaktoren auf den unterschiedlichen Ebenen im Hochschulsystem hat zu unterschiedlich ausgestalteten Prüfungssystemen und damit zu unterschiedlichen strukturellen Prüfungsbedingungen zwischen den Fächern, und innerhalb dieser auch zwischen den Hochschulen und zwischen Abschlüssen geführt (ebd.).

Diese Prüfungsbedingungen lassen sich für jede gewünschte Untersuchungseinheit ermitteln, da sie struktureller oder intersubjektiv geteilter Art sind. Die von Prahl als manifeste Normen bezeichneten Faktoren stellen formale Prüfungsbedingungen dar, die als solche anhand entsprechender Dokumente festgelegt sind. Sie umfassen bindende Regelungen, deren Wirksamkeit im Prüfungsprozess sich aus dem Zwang, sie anzuwenden ergibt. Dieser Zwang ist begründet durch die rechtliche Verbindlichkeit, unter anderem festgelegt durch das Hochschulrahmengesetz und durch die Hochschulgesetze der Länder. Da die Wirksamkeit formaler Regelungen im Prüfungsprozess rechtlich abgesichert ist, ist anzunehmen, dass sich Unterschiede zwischen den formalen Regelungen auch in unterschiedlichen Prüfungsergebnissen niederschlagen. Die als latente Normen bezeichneten Bedingungsfaktoren liegen nicht als formale Regelungen vor. Sie, wie auch die weiteren von Prahl genannten Elemente fachspezifischer Ausprägungen von Prüfungssystemen, stellen in der Regel jeweils für sich Produkte langfristiger Entwicklungen und Aushandlungsprozesse auf lokaler Ebene dar. Genau genommen sind sie damit alle örtliche und/oder fachliche Besonderheiten und ihre Ausprägungen sind abhängig von den jeweiligen Rahmenbedingungen, in denen die Prüfungen ablaufen, etwa von der üblichen Prü-

fungsbelastung, der Zusammensetzung der Lehrenden usw., in die die Benotungshandlungen der Prüfer*innen eingebettet sind.

Die langfristige Wirksamkeit dieser nicht-formalen akademischen Prüfungsbedingungen lässt sich kulturtheoretisch, und zwar anhand des Sozialisationskonzeptes und des Bourdieuschen Habitusbegriffs, erklären. Die Durchführung und Bewertung von Hochschulprüfungen muss wie jede andere soziale Handlung als sozial eingebetteter und, wie die Darstellung der Mechanismen der Benotung gezeigt hat, im bestimmtem Maße auch als interpretativer Akt verstanden werden (vgl. Shay 2005). Der interpretative, subjektive Einfluss wird zum einen durch die konkreten formalen Prüfungsbedingungen, denen die notengebende Einheit unterliegt, eingeschränkt. Zum anderen ist aber bereits die subjektive Komponente als solche keinesfalls ein reines Produkt individueller Präferenzen und Entscheidungen. Sie ist vorstrukturiert durch die im kollegialen bzw. disziplinären Rahmen herrschenden Präferenzen und Handlungsmuster (vgl. Wenger 1998), durch „frameworks, that are constituted in part as a result of membership within particular communities of practice“ (Shay 2005:665). Wer als Prüfer*in in einem bestimmten universitären oder fachspezifischen Kontext agiert, hat zuvor einen Prozess der fach- bzw. hochschulspezifischen Sozialisation durchlaufen, in dessen Verlauf die innerhalb der jeweiligen Gemeinschaft herrschenden Normen internalisiert werden (Gerholm 1990). Anders ausgedrückt bilden die wissenschaftlichen Gemeinschaften, innerhalb derer Prüfungs- und Benotungsprozesse durchgeführt werden, jede für sich eine Habitusklasse (vgl. Bourdieu 1998). Auch ohne weitreichende formale Regelungen sind Prüfende aufgrund ihrer habituellen Gemeinsamkeiten mit den nötigen Fähigkeiten und dem entsprechenden Wissen ausgestattet, um den Erwartungen der wissenschaftlichen Gemeinschaft gerecht zu werden (vgl. Shay 2005). Schon Studierende lernen als künftige Vertreter*innen akademischer Disziplinen, ihre eigenen Einstellungen und Handlungsmuster an die, die in ihrem Fach und an ihrer Hochschule vorherrschen, anzupassen. Ohne eine habituelle Übernahme der Denk- und Verhaltensweisen würden sie kaum mit dem akademischen Erfolg gesegnet werden, der sie zu Lehrenden und Prüfenden werden lässt. Mit der Aufnahme in den Kreis der Lehrenden wird das erlernte fach- und hochschulspezifische Wissen schrittweise um das Wissen der üblichen Handhabung bei der Prüfung und Benotung erweitert. Durch diese Ausbildung eines fach- und hochschulspezifischen Habitus wird den subjektiven Elementen der Notengebung ihre individuelle Beliebigkeit genommen und der Einfluss auf die im Prüfungs- und Benotungsverfahren zu treffenden Entscheidungen jenseits formaler Regelungen (mindestens) auf die nächst höher gelegene Ebene der an einer Hochschule tätigen Gemeinschaft einer (Organisationseinheit einer) fachlichen Disziplin verlagert.

Unterschiede zwischen „unter bestimmten historisch-gesellschaftlichen Rahmenbedingungen“ (Prahl 1983:439) entstandenen Prüfungssystemen äußern sich also in Form konkreter Zusammenspiele unterschiedlicher Prüfungsbedingungen: Sie umfassen die jeweils geltenden formalen Rahmenrege-

lungen als auch den jeweils internalisierten „Verhaltenscode“(Windolf 1992:77). Sowohl formale als auch nicht-formale Prüfungsbedingungen werden von den Prüfenden unmittelbar auf Prüfungs- und Benotungsprozesse angewendet. Sie (genauer: ihre Gesamtheit) entwickeln sich nur langsam und üben daher einen langfristigen, strukturellen Einfluss auf den Prozess der Notengebung aus. Die hier vertretene These ist, dass das Zusammenspiel konkreter Ausprägungen der im Prüfungs- und Bewertungsprozess enthaltenen Rahmenbedingungen dafür sorgt, dass die Noten einzelner akademischer Organisationseinheiten keine völlig willkürlichen Verzerrungen von abgefragter Leistung darstellen, sondern systematische Muster und eine gewisse zeitliche Stabilität aufweisen.

Wie gezeigt wurde, besteht der Notengebungsprozess aus mehreren Phasen mit unterschiedlichen Anschlussstellen für die Beeinflussung der Benotung. Unterschiede im Niveau der Noten zwischen Fächern bzw. Studiengängen können auf diversen leistungskonformen wie leistungsunabhängigen Ursachen beruhen. Theoretisch können sich die im Notengebungsprozess beinhalteten Prüfungsbedingungen einzelner Studiengänge in allen Ausprägungen voneinander unterscheiden – je mehr sie sich empirisch tatsächlich voneinander unterscheiden, umso wahrscheinlicher ist es, dass Differenzen im Notenniveau unabhängig von Differenzen im Leistungsniveau existieren. Im Folgenden werden potentiell prüfungsrelevante Distinktionsmerkmale zwischen Fächern (mit den klassischen Abschlüssen Diplom und Magister) herausgearbeitet.

3.2 Das Fundament der Fächerdifferenzierung: Fachkulturen

Fachkulturen stellen „einen Spezialfall kultureller Ausdifferenzierung“ (Windolf 1992:77) dar. Sie beschreiben die kulturelle Verfasstheit abgrenzbarer akademischer Organisationseinheiten. Der zugrunde liegende Kulturbegriff variiert dabei je nach theoretischer Ausarbeitung des Begriffes der Fachkultur, umfasst aber immer mindestens die Vorstellung geteilter „Wahrnehmungs-, Denk-, Wertungs- und Handlungsmuster“ (Liebau/Huber 1985:315). Anhand des Fachkulturenkonzepts wird die Erfassung von auf diese Weise als kulturell verstandenen Gemeinsamkeiten und Unterschieden zwischen akademischen Einheiten ermöglicht.

Das Konzept der Fachkulturen geht zurück auf die 1959 von Snow getroffene Gegenüberstellung von Naturwissenschaften und Geisteswissenschaften, betrachtet als zwei große Kulturen, die sich vor allem durch gegenseitiges Unverständnis des jeweils anderen „intellectual, moral and psychological climate“ (Snow 1959:2) voneinander abgrenzen. Diese erste analytische Differenzierung akademischer Einstellungen und Verhaltensweisen auf der Grundlage *kulturell*[10] verstandener Merkmale zog zahlreiche Diskussion um die kulturelle Beschaffenheit der akademischen Welt nach sich. Dabei wur-

[10] Die epistemologische Abgrenzung der Geistes- gegenüber den Naturwissenschaften vollzog sich bereits Ende des 19. Jahrhunderts und wird in der Regel mit dem Namen Wilhelm Dilthey verbunden (vgl. etwa Lessing 2001).

den immer tiefer gegliederte Ausdifferenzierungen akademischer Kulturen vorgenommen. Es erfolgten Erweiterungen der binären Unterscheidung Snows, wie bei Lepenies (1985) um den dritten kulturellen Block der Soziologie oder um die Gruppe der Gesellschaftswissenschaften und der angewandten Wissenschaften bei Gaff und Wilson (1971), über die Perspektive von Fachkulturen (Liebau/Huber 1985) bis hin zu Ansätzen, die bis in die Spezialisierungen einzelner Fächer hinein Möglichkeiten für die Ausbildung eigener Kulturen sehen (Huber 1990a).

Auf der anderen Seite wurde nicht nur von Kritikern der Annahme differenzierter akademischer Kulturmuster stets die Existenz einer übergeordneten akademischen Kultur mit gemeinsamen Werten betont (etwa Austin 1990). Erst allmählich setzte sich die Erkenntnis durch, dass eine allgemeine akademische Kultur und unterschiedliche, auf welchen Ebenen auch immer lokalisierte Fachkulturen sich nicht gegenseitig ausschließen, sondern vielmehr die Forschungsperspektive entscheidend für die Identifizierung von kulturellen Eigenheiten ist (vgl. Kuh/Whitt 1988; Multrus 2004)[11].

Der Vergleich kultureller Muster ist demnach zwischen empirisch identifizierbaren Einheiten auf jedem gewünschten Level möglich – von der Universitätsebene bis hin zu Spezialisierungen einzelner Fächer an einzelnen Hochschulen. Dabei muss jedoch klar sein, was genau eigentlich erforscht werden soll. Wie Multrus (2004) treffend feststellt, wird der Begriff Fachkultur in den meisten Studien, die sie zu untersuchen anstreben, nicht als eigenständiger Begriff definiert und als entsprechendes Konstrukt operationalisiert. Er wird stattdessen häufig als Synonym für Fächergruppen oder Fächer genutzt, die dann in der Analyse nicht nur durch ihre inhaltlichen Unterschiede, sondern zusätzlich „durch spezifische (kulturelle)Variablen zusätzlich unterschieden werden“ (Multrus 2004:179). Um Fachkultur als eigenständigen Begriff legitimieren zu können, nimmt Multrus eine empirische Identifizierung unterschiedlicher Fachkulturen vor, ohne vorher bereits Kategorisierungen nach Fächern oder Fächergruppen vorgenommen zu haben. In Folge seines clusteranalytischen Vorgehens fasst er Fachkulturen schließlich als „spezifische größere Gruppen einzelner Fächer, Studiengänge oder Fachrichtungen, deren Ähnlichkeiten innerhalb der Gruppen und deren Unterschiede zu anderen Fachkulturen sich insbesondere in kulturellen Merkmalen ausdrücken“ (ebd.:373) auf. Trotz dieses Verständnisses weist er in der Zusammenfassung seiner Untersuchung ausdrücklich darauf hin, dass sich akademische Organisationseinheiten auch auf anderen Analyseebenen sinnvoll anhand kultureller Merkmale unterscheiden und zuordnen lassen, eben auch anhand disziplinärer Abgrenzung (ebd.:398).

Neben der Frage, als welche Einheiten Fachkulturen gefasst werden, ist vor allem die Frage relevant, nach welchen Merkmalen sie unterschieden werden. In den letzten Jahrzehnten wurden zahlreiche empirische Studien zur Unterscheidung von Fächern nach kulturell verstandenen Eigenschaften

[11] Bei Multrus (2004) findet sich auch eine detaillierte Darstellung der Entwicklung der Forschung zu Fachkulturen, die deshalb an dieser Stelle nicht noch einmal ausführlich dargestellt wird.

durchgeführt. Die meisten dieser Untersuchungen, die üblicherweise als Beiträge zur Fachkulturforschung verstanden werden befassen sich entgegen der Multrusschen Auffassung des Begriffs Fachkultur mit Differenzen zwischen bereits inhaltlich unterscheidbaren Fächern oder Fächergruppen bzw. zwischen Fakultäten. Wissenschaftsorganisatorische und epistemologische Merkmale werden gelegentlich von den kulturellen Merkmalen eines Fachs unterschieden, nicht selten werden die erkenntnistheoretischen Eigenheiten der Fächer und ihre Organisationsstrukturen aber auch als Teil der spezifischen Fachkultur verstanden.
Arnold (2004) zeigt auf, dass alle fachkulturellen Unterschiede sich im Grunde auf wissenschaftsorganisatorische oder epistemologische Bedingungen zurückführen lassen, Fachkultur somit als Überbegriff für diese Unterschiede und die sich aus ihr ergebenden Denk-, Wahrnehmungs- und Handlungsmuster brauchbar ist, eine genuin kulturelle Komponente neben wissenschaftsorganisatorischen und epistemologischen Komponenten von Fachkulturen als eigenständige Kategorie zur Unterscheidung im Grunde jedoch redundant ist.

Viele Studien, die Unterschiede verschiedenster Art zwischen Fächern, Fächergruppen oder Fakultäten untersuchen und dabei prüfungsrelevante Merkmale beschreiben, wie beispielsweise Studierenden- oder Hochschullehrer*innenbefragungen weisen keinen Bezug zum Fachkulturkonzept auf. Nichtsdestotrotz geben sie Auskunft über Unterschiede zwischen Fächern, Fächergruppen und Fakultäten und werden deshalb, wenn informativ, im weiteren Verlauf in die Darstellung der Fachkulturforschung integriert.

3.2.1 Fachkulturen in der US-amerikanischen Hochschulforschung

Einer der ersten, der eine kulturelle Grenzziehung innerhalb der akademischen Welt vollzog, C.P. Snow, arbeitete dabei noch mit einem recht losen analytischen Rahmen. Die von ihm beschriebene Trennung von Geisteswissenschaften und Naturwissenschaften beruht vor allem auf seiner Wahrnehmung eines gegenseitigen Unverständnisses zwischen den beiden Lagern (Snow 1959). Unverständnis auf Seite der Geisteswissenschaften für die aus deren Sicht optimistische Grundhaltung der naturwissenschaftlichen Forschung, Unverständnis auf Seiten der Naturwissenschaftler*innen für eine angebliche Rückwärtsgewandtheit der Geisteswissenschaften. Snow versucht darzulegen, wie diese Missverständnisse aus seiner Sicht entstehen. Anschließend bemüht er sich, tatsächliche Unterschiede zwischen den beiden Seiten aufzuzeigen. Während Naturwissenschaftler*innen als zukunftsorientiert und hoch strukturiert, jedoch (auch im Privaten) ohne Verständnis für die Relevanz einer klassischen Kultur für Gegenwart und Zukunft dargestellt werden, werden Geisteswissenschaftler*innen als fixiert auf klassische Kultur ohne jeglichen Zugriff auf moderne wissenschaftliche Erkenntnisse fernab ihrer eigenen Perspektive gezeichnet.

Snows kulturelle Grenzziehung zwischen Akademiker*innen beruht auf einer Charakterisierung akademischer Organisationseinheiten als kulturelle Einheiten aufgrund seiner Wahrnehmung jeweils geteilter Einstellungen sowie gemeinsamer Arbeitsstandards und Denk- und Verhaltensweisen in den beiden Lagern. Die sieht er innerhalb der beiden Lager als unabhängig von Klassen- und Religionszugehörigkeiten und normativen Einstellungen wie etwa der politischen Ausrichtung. Auch wenn die von Snow wahrgenommenen Unterschiede zwischen Natur- und Geisteswissenschaften im weiteren Sinne als epistemologische Unterschiede beschrieben werden können, bleibt die Art der Unterscheidung vage, empirische Belege bleibt er, abgesehen von nicht weiter erläuterten Hinweisen auf selbst geführte Gespräche mit Vertreter*innen beider Lager, schuldig.

Während Snow als häufig zitierter Ideengeber für die analytische Trennung wissenschaftlicher Disziplinen und Fächer vor allem anstrebte, Grundcharakteristika der Geisteswissenschaften und Naturwissenschaften hervorzuheben, postulierte Gouldner bereits zwei Jahre vor Snows Veröffentlichung zu seiner These der zwei Kulturen ebenfalls eine dichotome Trennung akademischen Lebens, die dieses theoretisch wie empirisch wesentlich ausgereifter erfasst. Die Trennlinie zieht Gouldner dabei zwischen den beiden Gruppen der „Cosmopolitans“ und „Locals“ (Gouldner 1957:290). Diese generell auf die Mitglieder von Organisationen bezogene Unterscheidung vollzieht er anhand der drei Variablen Loyalität zur Organisation, Bindung an professionelle Fähigkeiten und Referenzgruppenorientierung. Den Cosmopolitans werden dabei eine geringe Loyalität gegenüber der sie beschäftigenden Organisation, eine hohe Bindung an ihre professionellen Fähigkeiten und eine Orientierung an Referenzgruppen außerhalb der eigenen Organisation zugeordnet. Die Locals zeichnen sich durch eine hohe Loyalität, eine niedrige Bindung an ihre speziellen Fähigkeiten und eine Orientierung an einer Referenzgruppe innerhalb der Organisation aus. Gouldner prüft seine These, dass sich die Mitglieder von Organisationen in die beiden Gruppen Locals und Cosmopolitans aufteilen lassen anhand von Interviews mit 125 akademischen Mitarbeiter*innen eines Kunst-Colleges, wobei er eine hohe Korrelation der als cosmopolitan bzw. der als local verstanden Ausprägungen der drei untersuchten Variablen als Beleg für die Existenz dieser Rollen versteht.
Anhand der Ausprägungen erstellt er einen Index, der die Einstufung als Local versus Cosmopolitan vierstufig darstellt und analysiert mit Hilfe dieses Indexes den Einfluss der beiden Gruppen auf 10 wichtige Entscheidungsbereiche und die Beteiligung an diesen innerhalb des Colleges. Es zeigt sich, dass sich diejenigen, die als stark cosmopolitan eingestuft werden im Einfluss kaum von denen unterscheiden, die als stark local eingestuft werden. Allerdings weisen diejenigen, die auf dem Index in die beiden mittleren Kategorien fallen, einen deutlich höheren Einfluss auf als die rein cosmopolitan und rein local Verorteten. Auch in der Beteiligung weisen die mittleren Kategorien die höchsten Werte auf, gefolgt von den Locals, die hier höhere Werte als die Cosmopolitans erreichen. Auch hinsichtlich

abgefragter Einstellungen zur Nutzung formeller Regeln zur Problemlösung zeigen die Locals höhere Zustimmung als die Cosmopolitans. Während diese Ergebnisse als solche keinen großen Beitrag zur Unterscheidung von Fächern zu leisten vermögen und wie Snows Arbeit auch keinen Bezug auf den Prüfungsprozess nehmen, sind sie, ebenso wie die These der zwei Kulturen als Fundament späterer Forschung von einiger Bedeutung. So ist die Unterscheidung zwischen Locals und Cosmopolitans von der Fachkulturenforschung vielfach aufgegriffen worden, es beruht etwa die Differenzierung von Fachkulturen nach ihrer kommunikativen Ausrichtung und der Reichweite ihrer Netzwerke auf ihr (vgl. etwa Becher 1987b; Huber 1990a; Köhler/Gapski 1997).

Unter anderem greift Clark 1963 Gouldners Unterscheidung zwischen Cosmopolitans und Locals auf und ergänzt sie um die Differenzierung „the pure versus the applied" (Clark 1963:42). Diese Trennung der reinen von den angewandten Wissenschaften beschreibt die vorwiegende Ausrichtung eines Faches auf den Erkenntnisgewinn entweder als Selbstzweck, also zur Weiterentwicklung der akademischen Disziplin oder als Basis für die praktische Anwendung. Clark sieht diese beiden Unterscheidungen als zwei wesentliche Analysekriterien zur Erfassung von kulturellen Typen in der akademischen Welt. Er erstellt anhand der Cosmopolitan/Local und pure/applied Unterscheidungen ein Vier-Felder Schema, aus dem vier idealtypische Professor*innentypen hervorgehen, die an Universitäten zu finden sind.

Den Vertreter*innen reiner Wissenschaft mit lokalem Bezug schreibt er dabei eine untergeordnete Stellung gegenüber den anderen drei Typen innerhalb des Lehrkörpers zu. Erwähnenswert ist Clarks Feststellung, dass die jeweilige Präferenz für eine kosmopolitische oder eine lokale Ausrichtung in direktem Zusammenhang mit der Lehrkultur steht: Eine lokale Ausrichtung fördert seiner Meinung nach den Kontakt zu Studierenden und geht mit einem höheren Stellenwert ihrer Ansichten und Probleme einher, als es bei einer kosmopolitischen Ausrichtung der Lehrenden der Fall ist (ebd. 1966). Trotz einem von ihm postulierten Trend zur kosmopolitischen Orientierung (ebd. 1963) betont Clark, dass die Hochschulen stets Platz für alle vier Typen der Orientierung bieten. Er fordert deshalb, dass um akademische Kultur verstehen zu können, sie nicht als einheitliches Gebilde gefasst werden dürfe, sondern Untersuchungseinheiten auch innerhalb von Hochschulen und innerhalb von disziplinären Grenzen gebildet werden, entsprechend der „departments and the array of disciplinary subcultures that (…) split the faculty" (ebd.:54).

Clark verweist dabei etwa auf Unterschiede im Zeitbudget, was das Verhältnis von Forschung und Lehre (auch Ruscio 1987: In Biologie und Physik werden mehr Wochenstunden auf Forschung verwendet als es etwa in den Sozialwissenschaften und der Politologie der Fall ist) sowie den Stellenwert von Forschung versus Lehre betrifft (In Physik und Biologie fällt ein geringeres Lehrdeputat an als in den Politikwissenschaften und der Anglistik und es findet sich, wie auch in Medizin und den Wirt-

schaftswissenschaften, eine Präferenz für die Forschung). Auch Unterschiede in der Organisation der Lehre in den einzelnen Instituten, die in Fächern mit einheitlichem Wissensfundus eher einvernehmlich, in Fächern mit konkurrierenden Ausrichtungen eher konflikthaft verläuft, werden von ihm genannt (Clark 1985 und 1987).

Mit Clark nimmt die Fachkulturforschung damit bereits in den 1960ern neben epistemologischen Merkmalen von Fächern auch die Lehrstrukturen an Hochschulen in den Fokus und entwickelt sich damit in nur wenigen Jahren von der Gegenüberstellung wahrgenommener Grundcharakteristika zweier Disziplinblöcke zur Analyse fachspezifischer Organisationsstrukturen und Lehrstile weiter.

Für die Analyse fachspezifischer Prüfungsbedingungen ergeben sich aus Clarks Arbeiten erste Ansatzpunkte: Ein unterschiedlich hohes Lehrdeputat und daraus resultierende Unterschiede im Zeitbudget für die Forschung sowie ein unterschiedlicher Stellenwert von Forschung könnten die Lehre und damit auch die Überprüfung des Gelehrten dahingehend beeinflussen, dass ein höheres Engagement sowie bessere Wissensvermittlung und gewissenhaftere Wissensabfrage bei einem Primat der Lehre bestehen. Bei einem Primat der Forschung könnten Prüfungen dagegen eher als lästiges Beiwerk gelten und möglicherweise zeitknapper geplant und durchgeführt werden. Ob ein Fach auf einem einheitlichen Wissensfundus aufbaut oder konkurrierende Paradigmen auf unterschiedlichen Grundlagen aufbauen, hat zudem möglicherweise einen Einfluss auf die Einheitlichkeit der Prüfungsgestaltung und damit auf die Vergleichbarkeit der Leistungen. Schließlich ist es denkbar, dass in reinen Wissenschaften Prüfungen eine andere Selektionsfunktion erfüllen als in angewandten - denn reine Wissenschaften bilden Studierende vorwiegend für den eigenen Nachwuchs aus und müssen daher auf ein hohes Niveau achten, um fähige Forschende zu produzieren und unter diesen nur den Besten auch Zutritt zum Wissenschaftsbetrieb zu ermöglichen. Die Absolvent*innen in den angewandten Wissenschaften hingegen werden vorwiegend für den externen Arbeitsmarkt ausgebildet und sind abhängiger vom dortigen Bedarf - möglicherweise entsteht damit dort ein variableres Selektionsklima als in reinen Wissenschaften.

Wie Clark zählen auch Gaff und Wilson zu den Kritikern der These von einer oder zwei großen akademischen Kulturen. Sie sind ebenfalls der Ansicht, dass mehrere „faculty cultures“ (Gaff/Wilson 1971:187) existieren und sich diese ‚Kulturen der Lehrenden‘ anhand der pädagogischen Werte, der Lehrorientierungen und Lebensstile der Professor*innen voneinander abgrenzen lassen. Sie untersuchen diese Annahme anhand postalischer Befragungen von sechs US-amerikanischen Colleges und Universitäten (ebd.).

Die Pädagogischen Werte werden über Meinungen zu diversen Items, die die Ziele einer College-Ausbildung, kontroverse Campus-Events und die Rolle von Studierenden in der Hochschulpolitik abbilden, abgefragt. Die Lehrorientierung wird anhand von Fragen über Lehrpraktiken und Ein-

stellungen zu Studierenden erfasst, der Lebensstil über die politische Einstellung, die religiöse Orientierung und Quellen der Lebensfreude. Die Ergebnisse, die Gaff und Wilson zunächst für die vier Gruppen Geisteswissenschaften, Naturwissenschaften, Gesellschaftswissenschaften und angewandte Wissenschaften (dazu zählen etwa Erziehungswissenschaften und Maschinenbau) darstellen, weisen für alle drei Kategorien wesentliche Unterschiede zwischen den Gruppen auf: Die Mitglieder der sozialwissenschaftlichen Fakultäten zeigen die liberalsten pädagogischen Werte, gefolgt von denen der Geisteswissenschaften. Professor*innen der Naturwissenschaften und angewandten Wissenschaften zeigen sich deutlich konservativer. Auch hinsichtlich politischer Einstellungen lassen sich die Vertreter*innen der Sozialwissenschaften am liberalsten, die der angewandten Wissenschaften am konservativsten einstufen (zur fachlichen Differenzierung politischer Einstellungen auch: Ladd/Lipset 1975). Es ist davon auszugehen, dass diese fachlichen Unterschiede in den (pädagogischen) Werten die Selektionsneigung von Prüfenden beeinflussen.

Gaff und Wilson differenzieren die vier Fakultätsgruppen weiter in einzelne Fächer und zeigen, dass auch zwischen ihnen erkennbare Unterschiede hinsichtlich der Merkmale Lehrpraktiken und Einstellungen gegenüber Studierenden herrschen: So zeigen sich die Mathematik Dozierenden in der Lehre am stärksten studierendenzentriert und weisen einen hoch strukturierten Unterrichtsstil auf, während den Studierenden in der Kunst und der Biologie der geringste Stellenwert bei der Ausrichtung der Lehrveranstaltungen beigemessen wird und in Geschichte und Philosophie der schwächste Strukturierungsgrad herrscht.

In einer weiteren Studie untersuchen Gaff et al. die Lernbedingungen von Studierenden einer niederländischen Universität in den Fächern Chemie, Jura, Psychologie und Medizin. Die Analyse der Ergebnisse zeigt fachspezifische Lernumgebungen und graduelle Unterschiede zwischen den Fächern hinsichtlich der Lernanforderungen an die Studierenden[12]. In Chemie wird von einem hohen Maß an benötigtem Faktenwissen, einer hohen Pflichtstundenzahl und einem hohen Strukturierungsgrad berichtet, während das Medizinstudium als ebenfalls durchstrukturiert, aber nicht im gleichen Maße fordernd wahrgenommen wird. Die Lernvorgaben in Jura und Psychologie werden im Gegensatz zu diesen Fächern eher als unbestimmt wahrgenommen. Der zum Lernen benötigte Arbeits- und Zeitaufwand und die Erwartungen an das Engagement der Studierenden sind entsprechend in Chemie am höchsten, in Psychologie und Jura am geringsten (Gaff et al. 1977).

Nun gehen nicht mehr nur die Lehrenden und die Lehrstrukturen der Fächer in die Analyse der Fachkulturen ein, auch ihre Auswirkungen auf Studierende und deren Lernbedingungen werden betrachtet, was eine Integration aller wissenschaftlich arbeitenden Beteiligten an der Hochschule in die Betrachtung bedeutet. Die Fachkulturforschung hat damit den Bogen von den epistemologischen Grundlagen (Snow, Clark) bis zum wissenschaftlichen Nachwuchs gespannt und damit ihr Untersu-

[12] Zu unterschiedlichen Lernumgebungen siehe auch Winteler (1981).

chungsfeld ausgeweitet. Dabei ist es naheliegend anzunehmen, dass fachspezifische Lernumgebungen als Folge der jeweiligen Strukturen der Wissensvermittlung und –abfrage entstehen, welche wiederum auf den epistemologischen Charakteristika einer Disziplin beruhen, und damit eine Reaktion auf fachspezifische Prüfungsstrukturen darstellen. Je standardisierter die Prüfungsinhalte, umso eher werden auch die Lernanforderungen strukturierte Wissensvermittlung denn interpretative Fähigkeiten umfassen. Es kann daraus gefolgert werden, dass aufgrund der disziplinspezifischen Erkenntnisweise hoch strukturierte und hoch standardisierte Lehre und Lernumgebungen auch in entsprechend standardisierte Prüfungen münden, während umgekehrt ein geringer Standardisierungsgrad in den Inhalten und deren Vermittlung auch kaum zu hoch standardisierten Abfragen dieser Inhalte führen dürfte.

Biglan (1973) erfasst diese Differenz im Standardisierungsgrad in ihrem Ursprung und hebt die epistemologische Unterscheidung nach der Existenz eines einheitlichen Paradigmas hervor. Über die Befragung von Hochschullehrenden einer US-amerikanischen Universität und eines US-amerikanischen Colleges gelangt er zu einer der prominentesten Unterscheidungslinien zwischen Fächern. Er kombiniert die Unterscheidung nach der Existenz eines einheitlichen Paradigmas mit der nach Orientierung auf Wissenschaft oder auf Praxis (wie auch Clark 1962) und mit der nach Beschäftigung mit lebenden oder leblosen Inhalten.
Diese Kombinationen ergeben jeweils drei Vier-Felder-Kategorien. Während die Unterscheidung nach lebendigen und leblosen Inhalten in der folgenden Forschung keine nennenswerte Beachtung findet, stellen die Unterscheidungen nach harten (ein einheitliches Paradigma verwendend) und weichen (kein einheitliches Paradigma verwendend), sowie nach angewandten (praxisorientiert) und reinen (wissenschaftsorientiert) Fächern vielgenutzte Analysedimensionen zur Charakterisierung und Kategorisierung von Fächern dar. Biglan gelangt anhand der Kombination der Dimensionen hart-weich und rein-angewandt zu einer Charakterisierung von Physik, Biologie, Chemie, Geologie, Astronomie und Mathematik als hart/rein und von Maschinenbau, Agrarwissenschaften und Informatik als hart/angewandt. Erziehungswissenschaftliche Studiengänge, Kommunikationswissenschaften und betriebswirtschaftliche Fächer zählt er zur Kategorie weich/angewandt, und Psychologie, Philosophie, Geschichte, Soziologie, Politikwissenschaften, Anglistik und Germanistik werden als weich/rein eingestuft (Biglan 1973).
Da die Unterscheidung nach weichen und harten Fächern auf dem Grad der Standardisierung in der Erklärung untersuchter Phänomen aufbaut, ist in ihr der Ausgangspunkt für die bei Gaff und Wilson beobachteten Differenzen in der Strukturierung und Standardisierung von Lehr- und Lernumgebungen zu finden. Entsprechend dürfte sich, wie bereits zuvor beschrieben, in den Prüfungen eine Differenz zwischen eher interpretatorischen Leistungen in den weichen Fächern und eher formalisierten

Leistungen in den harten Fächern wiederfinden. Wird das Vier-Felder Schema auf die Prüfungssysteme der Fächer übertragen, ergibt sich eine idealtypische Einteilung in Fächer mit a) vorwiegend interpretatorischen Prüfungsleistungen und einem konstanten Selektionsniveau, b) vorwiegend interpretatorischen Prüfungsleistungen und einem variablen Selektionsniveau, c) vorwiegend formalisierten Prüfungsleistungen und einem konstanten Selektionsniveau und d) vorwiegend formalisierten Prüfungsleistungen und einem variablen Selektionsniveau.

Whitley widmet sich 1984 den Organisationstrukturen des modernen Wissenschaftsbetriebs. Er beschreibt moderne Wissenschaft als „system for co-ordinating and controlling the production of innovations as a collective enterprise" (Whitley 1984:85). Die wissenschaftliche Arbeit ist eingebettet in konkrete Rahmenbedingungen der Kooperation, die von Fach zu Fach hinsichtlich der Art, in der neues Wissen generiert und der Beitrag von neuem Wissen gestaltet wird, variieren. Der Beitrag von neuem Wissen, der gleichzeitig die Basis wissenschaftlicher Reputation darstellt, muss nach Whitley zwei zentrale Kriterien erfüllen: Einerseits müssen tatsächlich neue Erkenntnisse produziert werden, Wissen also, das sich von dem bisherigen Wissen innerhalb der Scientific Community unterscheidet, andererseits muss dieses Wissen nach geteilten Maßstäben produziert werden und potentiell von anderen Forschenden genutzt werden können, um deren eigene Arbeit voranzubringen. Wissenschaftliches Arbeiten ist daher geprägt durch die Abhängigkeit von anderen Forschenden und durch eine gewisse Aufgabenunsicherheit, so Whitley.

Der Grad der Abhängigkeit Forschender von anderen Forschenden bildet deren Relevanz für das eigene wissenschaftliche Ansehen ab. Er wird vor allem durch das Ausmaß, in dem Forschungsergebnisse zu den Ergebnissen und dem Vorgehen von Kolleg*innen passen müssen, um deren Anerkennung zu erhalten und in dem es möglich ist, eigene Arbeiten unterschiedlichem Publikum zu präsentieren, bestimmt. Der Grad der Aufgabenunsicherheit ist ein Maß der Möglichkeiten, Forschungsergebnisse vorherzusagen, sichtbar zu machen und zu wiederholen. Je geringer diese Möglichkeiten sich aufgrund unterschiedlicher Vorgehensweisen und Vorstellungen über wissenschaftliche Standards darstellen, umso höher ist die Aufgabenunsicherheit.

Im Bestreben, einerseits neues Wissen zu generieren, andererseits die Erwartungen der Scientific Community zu erfüllen, sieht Whitley ein Konfliktpotential, das zu einem von Fach zu Fach unterschiedlich ausgestaltetem Balanceakt zwischen Abhängigkeit und Aufgabenunsicherheit führt. Diese konkrete Konfiguration des Ausgleichs ist seiner Meinung nach für fachspezifische Unterschiede in der Wissensproduktion verantwortlich.

Die gegenseitige Abhängigkeit lässt sich in eine als funktional und eine als strategisch bezeichnete Komponente unterteilen. Funktionale Abhängigkeit beschreibt das Ausmaß, in dem Wissenschaftler*innen „specific results, ideas, and procedures of fellow specialists" (ebd.:88) adaptieren müssen,

um Anerkennung für ihre Arbeit zu erhalten. Sie bezieht sich damit vorwiegend auf den Erkenntnisgewinn und ist als epistemologisches Merkmal zu verstehen – je höher der Grad an funktionaler Abhängigkeit, umso limitierter sind die Möglichkeiten, Forschung unabhängig von bestimmten ‚best practice' Vorgehen zu betreiben. Strategische Abhängigkeit lässt sich als Notwendigkeit, die Mitglieder der wissenschaftlichen Gemeinschaft von der Bedeutung der untersuchten Problemstellung sowie des eigenen Vorgehens für die gemeinsame Wissenskumulierung zu überzeugen, um deren Anerkennung zu erhalten und stellt damit ein rein wissenschaftsorganisatorisches Charakteristikum dar.
Auch die Aufgabenunsicherheit lässt sich in zwei Komponenten gliedern, eine technische und eine strategische. Technische Aufgabenunsicherheit meint dabei Unklarheiten über die Anwendung von Methoden und die Interpretation von Ergebnissen, die die Vorhersage, Sichtbarkeit und Wiederholbarkeit von Forschungsergebnissen einschränken. Strategische Unsicherheit bezieht sich auf Uneinigkeiten innerhalb einer Disziplin über die Hierarchisierung von Problemstellungen und Vorgehensweisen, die keine Aussagen über die Relevanz von Forschungsergebnissen im Vorfeld einer Untersuchung zulassen. Der Grad der technischen Aufgabenunsicherheit beeinflusst wie der der funktionalen Abhängigkeit vor allem die Erkenntnisproduktion. Je höher der Grad an technischer Aufgabenunsicherheit, umso unsicherer ist, ob die Ergebnisse der Forschung tatsächlich neues Wissen darstellen. Die strategische Komponente der Aufgabenunsicherheit hingegen ist wie die strategische Abhängigkeit ein wissenschaftsorganisatorisches Merkmal - auch hier steht nicht im Vordergrund, wie das neue Wissen zustande kommt, sondern wie schwierig es ist, die scientific community davon zu überzeugen, dass das verwendete Vorgehen für das Fach Relevanz besitzt.
Beeinflusst werden der Grad der Abhängigkeit und der Aufgabenunsicherheit durch das jeweilige Ausmaß der Autonomie in der Reputationssetzung, die jeweilige Konzentration von Kontrolle über die Mittel der Produktion und Distribution von Wissen und die Pluralität und Diversität des Publikums. Die Autonomie in der Reputationssetzung und damit auch die gegenseitige Abhängigkeit sinken mit zunehmender Vielfalt an Methoden, die geeignet sind, valides Wissen zu produzieren. Je diverser die Gruppe und die Methodensammlung ist, mit der Wissen produziert werden kann, umso höher wird außerdem die Aufgabenunsicherheit. Die Kontrolle über die Mittel der Produktion und Verbreitung von Wissen ist abhängig von der Verteilung materieller Ressourcen und dem Zugang zu Kommunikationsstrukturen, anhand welcher Reputation erlangt werden kann. Je breiter verteilt materielle Ressourcen sind und je einfacher der Zugang zu relevanten Kommunikationskanälen ist, umso schwieriger gestaltet sich die Kontrolle über das, was an Wissen produziert wird und dessen Diffusion, was Abhängigkeiten ebenfalls abbaut. Geringe Kontrolle über materielle Ressourcen fördert allerdings strategische Aufgabenunsicherheit, da diese Ressourcen dann verschiedenen Gruppen mit unterschiedlichen Zielen zur Verfügung stehen. Schließlich vermindert ein breit gefächertes und möglichst uneinheitliches Publikum für Forschungsergebnisse gegenseitige Abhängigkeit, da die Umset-

zung einer breiten Palette von Forschungszielen und -methoden gefördert wird. Dadurch wird jedoch wiederum die Aufgabenunsicherheit erhöht.

Whitley vollzieht mehrere Kombinationen der von ihm generierten Dimensionen und deren Ausprägungen. So erstellt er zwei Vier-Felder-Schemata, die für die jeweiligen Komponenten der Abhängigkeit und der Aufgabenunsicherheit die Ausprägungen hoch und niedrig enthalten und ordnet diesen wissenschaftsorganisatorische Merkmale und Beispielfächer zu. Fächer mit hoher strategischer und niedriger funktionaler Abhängigkeit sind gekennzeichnet durch klar voneinander abgegrenzte Forschungsschulen, die unterschiedliche Forschungsziele mit unterschiedlichen Vorgehensweisen und Methoden verfolgen. Innerhalb der einzelnen Schulen herrscht ein hoher Grad an Kooperation, zwischen ihnen allerdings starke Konkurrenz um die Vorherrschaft im Fach. Als Beispiel werden die Philosophie und Psychologie Deutschlands vor 1933 angeführt. Im Falle niedriger strategischer und niedriger funktionaler Abhängigkeit ist neben der Vielzahl parallel existierender Forschungsziele und Methoden auch eine Vielzahl an nur lose miteinander verbundenen Forschungsgemeinschaften zu beobachten, unter denen kaum wissenschaftliche Koordination stattfindet, wie Whitley es für die Soziologie diagnostiziert. Fächer mit geringer strategischer und hoher funktionaler Abhängigkeit setzen sich aus spezialisierten Forschungsgemeinschaften mit spezifischen Zielen zusammen, die innerhalb ihrer Spezialisierung standardisierte Methoden anwenden. Beispiele sind die Chemie und die US-amerikanische Mathematik. Sind beide Komponenten der Abhängigkeit hoch, zeichnen sich die Fächer neben dem hohen Spezialisierungsgrad außerdem durch eine hohe Koordination der Problemstellungen und Forschungsziele aus. Innerhalb des Fachs herrscht ein Wettbewerb um die hierarchische Gliederung der einzelnen Fachrichtungen, wie es in der Physik der Fall ist.

Eine hohe strategische Aufgabenunsicherheit in Kombination mit einer niedrigen technischen Aufgabenunsicherheit geht einher mit vorhersagbaren, transparenten und wiederholbaren Forschungsergebnissen und hoher Übereinstimmung, wie diese Ergebnisse zu interpretieren sind. Über Problemstellungen und Ziele herrscht allerdings keine einheitliche Meinung. Beispiele hierfür sind etwa die Biologie und die Ingenieurswissenschaften. Niedrige strategische und hohe technische Aufgabenunsicherheit steht umgekehrt für eine klare Strukturierung der Ziele und Problemstellungen bei gleichzeitigen Problemen, zu reliablen Ergebnissen und intersubjektiven Interpretationen zu gelangen. Als Beispiele führt Whitley die Wirtschaftswissenschaften an. Sind beide Komponenten nur in geringem Maße vorhanden, herrscht neben der Übereinstimmung hinsichtlich der Anwendung von Methoden und der Interpretation von Ergebnissen auch Einigkeit über die Rangfolge von Problemstellungen und kollektiven Zielen, so wie in Physik und Chemie. Ein hoher Grad in beiden Unsicherheitstypen bedeutet Unklarheiten hinsichtlich der Methodenanwendung und Interpretation von Ergebnissen bei gleichzeitiger Uneinigkeit hinsichtlich relevanter Forschungsziele und -problemen. Beispielhaft wird für diese Kombinationsmöglichkeit auf die US-amerikanische Soziologie verwiesen.

Neben den Vier-Felder-Tafeln, die die möglichen Kombinationen der Dimensionen Abhängigkeit und Aufgabenunsicherheit abbilden, beschreibt Whitley außerdem die möglichen Verbindungen von Ausprägungen der beiden Dimensionen zwischen einander. Aufgrund der Einschätzung, dass einige der theoretisch möglichen 16 Kombinationen in der Praxis unwahrscheinlich seien bzw. nur unter sehr speziellen Umständen bestehen könnten, reduziert er die Kombinationsmöglichkeiten schließlich auf sieben Typen von Organisationsstrukturen wissenschaftlicher Tätigkeit.

Sie ergeben sich aus zwei Vier-Felder-Kombinationen: Zum einen aus der Gegenüberstellung von geringer technischer Aufgabenunsicherheit bei gleichzeitiger a) hoher und b) niedriger strategischer Aufgabenunsicherheit auf der einen Seite und hoher funktionaler Abhängigkeit bei gleichzeitiger a) hoher und b) niedriger strategischer Abhängigkeit. Zum anderen aus der Gegenüberstellung von hoher technischer Aufgabenunsicherheit bei gleichzeitiger c) hoher und d) niedriger strategischer Aufgabenunsicherheit auf der einen Seite und niedriger funktionaler Abhängigkeit bei gleichzeitiger c) hoher und d) niedriger strategischer Abhängigkeit, wobei die Kombination d)+d) als instabil betrachtet wird und wegfällt (ebd.).

Die strategischen Komponenten von Abhängigkeit und Aufgabenunsicherheit beziehen sich als wissenschaftsorganisatorische Merkmale vor allem auf den zu leistenden Aufwand für zukünftige Forschung im Fach und weisen damit keinen Einfluss auf den bereits bestehenden Wissenskanon auf, der den Prüfungsinhalt darstellt. Anders ist es bei technischer Aufgabenunsicherheit und funktionaler Abhängigkeit, die als epistemologische Merkmale einzuordnen sind. So gehen ein niedriger Grad an technischer Aufgabenunsicherheit und hohe funktionale Abhängigkeit vermutlich mit einer relativ hohen Vergleichbarkeit von Prüfungsaufgaben einher. Denn wenn es üblich ist, die Forschung auf zentrale Erkenntnisse und Vorgehensweisen und auf eine hoch standardisierte Methodenanwendung aufzubauen, wird diese standardisierte Arbeitsweise wohl auch schon in Prüfungen vermittelt. Umgekehrt ist bei einem hohen Grad an technischer Aufgabenunsicherheit davon auszugehen, dass sich die Differenzen in der Wissensproduktion und -beurteilung einschränkend auf die Vergleichbarkeit von Prüfungsaufgaben und -leistungen etwa zwischen mehreren Hochschulen oder im Extremfall sogar zwischen einzelnen Prüfer*innen innerhalb eines Instituts auswirken. Ein geringer Grad an funktionaler Abhängigkeit, in diesem Punkt vergleichbar mit der Nichtexistenz eines einheitlichen Wissensfundus, dürfte ebenfalls eher zu einer inhaltlichen Heterogenität im Prüfungswesen führen, da hier die Möglichkeiten konkurrierender Schulen ihre jeweiligen Paradigmen durchzusetzen und in der Lehre zu reproduzieren größer sind als bei einem hohen Maß an Abhängigkeit.

Becher (1981) identifiziert Unterschiede in den sechs Fächern Physik, Geschichte, Biologie, Soziologie, Maschinenbau und Jura anhand von 126 Interviews mit Lehrenden und Forschenden jeweils dreier bzw. vierer (Jura und Soziologie) US-amerikanischer Universitäten. Er verzichtet auf eine Ge-

genüberstellung der einzelnen Fächer mittels der Aussagen der Befragten. Stattdessen konzentriert er sich auf die seines Erachtens wesentlichen Unterschiede und Gemeinsamkeiten zwischen den Fächern. Er hebt dabei Unterschiede in der gegenseitigen Wahrnehmung zwischen den Fächern (zusammenfassend Huber 1990a:87) hervor. Aber auch epistemologische und kulturelle Unterschiede werden vermittelt. Während in den Geschichtswissenschaften und in der Biologie etwa die Komplexität der Wirklichkeit betont wird, teilen Vertreter*innen der Physik und der Ingenieurswissenschaften die Ansicht, letztendlich könne alle Komplexität auf einfache Bausteine heruntergebrochen werden. Der Forschungsstil ersterer wird eher als empirisch geleitet, der in der Physik und in der Soziologie als theoriegeleitet beschrieben. In Physik, Biologie und Maschinenbau wird die Rolle von ideologischen Motiven als unbedeutend, in Geschichte und Soziologie als eher einflussreich beschrieben. Auch Unterschiede im Grad der inhaltlichen und methodischen Abgrenzung zu anderen Fächern (in Geschichte etwa sehr niedrig), im Publikationsverhalten (kurze Artikel in der Physik, Journals im Maschinenbau, Monographien und Journals in der Biologie und in Jura, der höchste Wert für Bücher in Geschichte und für Beiträge in Journals in der Soziologie, Notwendigkeit der Zusammenfassung des Forschungsstandes in der Soziologie und in Geschichte) werden aufgezeigt (Becher 1981; 1984).

Becher betont dabei, dass auch innerhalb der einzelnen Fächer relevante Unterschiede zwischen verschiedenen Organisationseinheiten, etwa zwischen theoretisch und experimentell ausgerichteten Gruppen bestehen können, die es erfordern, die Untersuchung von Fachkulturen tiefer als nur auf der Fachebene zu betreiben (Becher 1981; 1990). Bechers Analysen zeigen, dass die fachliche Sozialisation ein Prozess ist, der sich in vielen einzelnen Bereichen des wissenschaftlichen Arbeitens vollzieht, deren Spezifika nach und nach erkannt und internalisiert werden. Von Fach zu Fach werden dabei schon im Studium beginnend verschiedene Lektionen fachlicher Besonderheiten im Wissenschaftsbetrieb offen unterrichtet oder implizit weitergegeben. Auf diese Weise entstehen die fachspezifischen Erkenntnis- und Arbeitsweisen, die Becher auf fachspezifische Verhältnisse und Zugänge zum Wissen zurückzuführt.

Diese unterschiedlichen epistemologischen Fundamente der Fächer führen, wie Becher anhand von vorhandener Forschungsliteratur und eigenen Analysen von Fachjournalen zeigt, unter anderem zu Unterschieden im Vokabular, im implizitem Wissen (siehe hierzu auch Gerholm 1990; Arnold 2004), im Aufbau der Veröffentlichungen und in deren Länge sowie in der Struktur und dem Aufbau von Argumenten: In der Soziologie erkennt Becher einen Fokus impliziten Wissens auf methodische und theoretische Verfahrensweisen, in Physik charakterisiert er die impliziten Wissensbestände dagegen als auf Einstellungen und Verhaltensweisen ausgerichtet. Soziolog*innen und Historiker*innen bewerten in Reviews auch die Relevanz von Beiträgen ihrer Kolleg*innen zur Forschung, Physiker*innen bewerten die berichteten Fakten, wobei die Sprache, in der Forschungsergebnisse dargestellt wer-

den, sich bei den Historiker*innen der Umgangssprache nähert, in der Physik mit ihren zahlreichen Codierungen als Fachsprache bezeichnet werden kann. Journalbeiträge von Physiker*innen sind deutlich kürzer als die in der Soziologie, Historiker*innen legen im Schnitt die längsten Artikel vor. Physiker*innen, als Vertreter*innen reiner Wissenschaft verfolgen eine atomistische, kumulierende Wissensbildung und formen ihre Argumentationslinien entsprechend dieser Erkenntnisweise aus aufeinander aufbauenden Komponenten, deren Struktur im Voraus feststeht. In der Soziologie und in Geschichte hingegen wird die Argumentationsstruktur jeweils dem spezifischen Thema angepasst, so Becher. Die konkrete Festlegung der Argumentationslinie basiert hier unter anderem auf der Darstellung des bisherigen Forschungsstandes, dessen Zusammenfassung in Publikationen der Physik nicht unbedingt nötig ist (Becher 1987a).

Was Becher mit seinen Arbeiten gelingt, ist, den Zusammenhang zwischen den zuvor auch schon von anderen als kulturell definierten fachlichen Besonderheiten in der Wahrnehmung, im Denken und im Handeln und den epistemologischen sowie wissenschaftsorganisatorischen Strukturen der Fächer aufzuzeigen: 1987 systematisiert Becher die sechs 1981 untersuchten Fächer, indem er sie einer der vier Ausprägungen, die sich aus der Kombination der Kriterienpaare hart/weich und rein/angewandt ergeben, zuordnet. Anschließend beschreibt er für jede der auf diese Weise gebildeten Fächergruppen (hart-rein; hart-angewandt; weich-rein; weich-angewandt) die seinen Analysen zufolge entsprechend zugehörigen kulturellen Verhaltensweisen und epistemologischen Merkmale (welche sich in die beiden Unterkategorien Erkenntnisweise und Erkenntnisziel aufteilen lassen, so geschehen bei Huber, 1990a:79). Diese Beschreibungen basieren auf Erkenntnissen aus seinen bereits 1981 vorgestellten Interviews, die er nun entlang der Themenkomplexe Einführung in die akademische Profession, soziale Interaktionen, Spezialisierung sowie Wandel und Mobilität gliedert (Becher 1987b).

Bis 1989 erweitert Becher seine bisherigen Untersuchungen zu Fachkulturen auf insgesamt 12 Fächer, zu denen er nun über Daten aus 18 Hochschuleinrichtungen in den USA und Großbritannien verfügt. In „Academic Tribes and Territories“ präsentiert er die Schlussfolgerungen, die er aus seinen Analysen des Datenmaterials hinsichtlich der Unterschiede zwischen Fachkulturen zeiht. Becher stellt hier den Zusammenhang zwischen den kulturellen Denk- und Verhaltensweisen der einzelnen Fächer und deren jeweiligen epistemologischen Merkmalen, den er bereits 1987 erkannt hat, in den Vordergrund: „It would seem, then, that the attitudes activities and cognitive styles of groups of academics representing a particular discipline are closely bound up with the characteristics and structures of the knowledge domains with which such groups are professionally concerned” (Becher 1989:20).

Becher fasst in „Academic Tribes and Territories“ vornehmlich die Ergebnisse seiner bereits in den Jahren zuvor veröffentlichten Untersuchungen zusammen und führt sie detaillierter als in den Forschungsartikeln aus. Er hebt neben den Klassifizierungsdimensionen hart-weich und rein-angewandt Unterscheidungen zwischen eng verwobenen (konvergenten) und lose miteinander verknüpften (di-

vergenten) Wissensgemeinschaften (in Bezug auf Ideologien, Qualitätskriterien, geteilte Normen und Werte, Netzwerke etc.), zwischen ländlichen und urbanen Forschungsstilen und innerhalb der Kommunikationsmuster als geeignete Analysekategorien zur Unterscheidung von Fachkulturen vor. Er bezieht sich dabei zum großen Teil auf die bereits 1981 von ihm eingeführte Unterscheidung nach ländlichem versus urbanem Forschungsstil.

Als urbanen Forschungsstil beschreibt er die Aufteilung eines Problems in mehrere Komponenten, die jede einzeln bearbeitet wird (atomistisches Vorgehen), rural ist demzufolge ein holistischer Ansatz, der die unabhängige Behandlung einzelner Teile eines Phänomens als Ignoranz gegenüber dessen Komplexität betrachtet. Dementsprechend geht ein urbaner Forschungsstil mit kurzer Forschungsdauer und begrenzten Ergebnissen einher und wird öfter in (teils kompetitiver) Kooperation durchgeführt, während rurale Forschende lange Zeit mit komplexen Fragen beschäftigt sind und arbeitsteilig vorgehen, um umfassende Ergebnisse zu erzielen. Die Kommunikationsstrukturen urbaner Forscher*innen sind kosmopolitisch, also lose und weitreichend, rurale Netzwerke sind eng und von begrenztem Ausmaß (ebd. 1981). Enge Verknüpfungen innerhalb der Gemeinschaften gehen für Becher mit klaren Abgrenzungen dieser Gemeinschaften einher, während lose, weitläufige Gruppen weniger klare Abgrenzungen gegenüber Vertreter*innen anderer Fächer vornehmen (ebd. 1989). Erstere weisen stabile und einheitliche, Letztere eher instabile, uneinheitliche epistemologische Merkmale auf.

In Ergänzung seiner vorangegangenen Arbeiten verweist Becher zudem auf die Relevanz der Beziehungen der einzelnen Fächer zu den übrigen gesellschaftlichen Bereichen und ihres gesellschaftlichen Status für die Analyse von Fachkulturen. Demnach weisen vor allem die Fächer der Kategorien hart/rein und hart/angewandt hohes Prestige in der Gesellschaft auf, sind allerdings auch in hohem Maße abhängig vom öffentlichen und privaten Wirtschaftssektor. Fächer, die als weich/rein eingestuft werden können, weisen eine geringere externe Abhängigkeit auf, genießen außerhalb der akademischen Welt allerdings auch ein geringeres Ansehen als harte Fächer. Dies gilt auch für Fächer der Kategorie weich/angewandt, die, so Becher, jedoch auch noch stark abhängig von den Interessenverbänden der jeweiligen Professionen sind (ebd.).

Bechers finaler Schluss postuliert schließlich eine Übereinstimmung der vier Unterscheidungsfelder nach den epistemologischen Kriterien hart-weich und rein-angewandt mit den vier sozialen Unterscheidungsfeldern nach den Kriterien konvergent/divergent und ländlich/urban (ebd. und als Abbildung: Huber 1990a:79). Er klassifiziert Physik, Chemie, Pharmazie, Mathematik und (mit Abstrichen) Biologie und Maschinenbau als harte Fächer, Geschichte, Sprachwissenschaften, Soziologie und Jura als weiche Fächer, Wirtschaftswissenschaften und Geographie als Grenzfälle. Von diesen Fächern stuft er Maschinenbau, Jura und Pharmazie als angewandte Fächer ein. Geschichte, Wirtschaftswissenschaften, Mathematik und Physik stellen für ihn Fächer mit konvergenten Sozialformen dar, Bio-

logie, Chemie und Jura markieren den Übergang zu den übrigen, als divergent eingestuften Fächern. Lediglich für Physik konstatiert er einen urbanen Forschungsstil (ebd.).

Becher hat damit die umfassendste Analyse von fachkulturellen Besonderheiten in der US-amerikanischen Forschung vorgelegt, die neben der Ausarbeitung konkreter Ausprägungen vor allem eine Integration zuvor in anderen Studien separat betrachteter Analysefelder zum Ergebnis hat. Die meisten bei Becher genauer dargestellten Distinktionskriterien stehen aus diesem Grund in Zusammenhang mit den bereits besprochenen prüfungsrelevanten Merkmalen. So dürfte in diesem Verständnis etwa die Unterscheidung nach einem komplizierenden vs. simplifizierenden Wirklichkeitsverständnis mit dem Grad der Standardisierung einhergehen - je simplifizierender das Verständnis, umso standardisierter lässt es sich abfragen. Die Annahme, die erforschte Wirklichkeit lässt sich auf fundamentale Bausteine zurückzuführen lässt zudem einen einheitlicheren Wissensfundus erwarten als die Einstellung, dass die erforschte Welt mehr als die Summe ihrer Teile darstellt. Beide Zusammenhänge zeigen sich im Publikationsverhalten - die Betonung von Komplexität geht mit Erklärungsmehraufwand einher und damit mit steigender Heterogenität. An dieser Stelle ist auch der Bezug zu Withley zu erkennen, da eine simplifizierende Erklärungsweise einen geringen Grad an technischer Aufgabenunsicherheit und einen hohen Grad an funktionaler Abhängigkeit erfordert – eine Kombination, die wiederum hohe Formalisierung und Standardisierung im Prüfungsprozess erwarten lässt.

Neu ist bei Becher die Zuschreibung der Ideologierelevanz im wissenschaftlichen Betrieb. Wird die Unterscheidung nach objektiv, ‚nüchtern' forschenden gegenüber in ihrer Interpretation ideologisch vorbelasteten Wissenschaftler*innen auf den Beurteilungsprozess von Prüfungen übertragen, ließe sich vorstellen, dass letztere Gruppe gegenüber äußeren Einflüssen auf die Beurteilung empfänglicher reagiert und damit im Zeitverlauf instabilere Selektionsneigungen aufweist. Deutlicher als bisher betont Becher zudem die Verbindungen eines Fachs zur Restgesellschaft außerhalb der Hochschule. Reputation und Abhängigkeit von den übrigen Gesellschaftsbereichen, allen voran vom Wirtschaftssektor, sollten vornehmlich die Finanzierungsstrukturen beeinflussen. Gerade bei outputbedingter Förderung ist jedoch ein Effekt auf die Beurteilungspraxis denkbar.

Wie die Übersicht über die zentralen Arbeiten der US-amerikanischen Forschung im Bereich Fachkulturen zeigt, lassen sich Fächer neben der inhaltlichen Orientierung nach unterschiedlichen als epistemologisch, wissenschaftsorganisatorisch und kulturell verstandenen Merkmalen unterscheiden. In Hinblick auf den Prüfungsprozess lassen sich verschiedene Distinktionskriterien erfassen, die die Ausgestaltung fachspezifischer Prüfungssysteme begünstigen. Sie umfassen

- die Existenz eines einheitlichen Wissensfundus
- die Existenz eines einheitlichen Paradigmas

- das Wirklichkeitsverständnis: komplizierend vs. simplifizierend
- den Grad der funktionalen Abhängigkeit und der technischen Aufgabenunsicherheit
- die Bedeutung von Ideologien
- den gesellschaftlichen Status und die Beziehungen zur Restgesellschaft
- den Grad der Standardisierung in Lehr- und Lernstrukturen
- den Lehrstil und die damit einhergehende Kontakthäufigkeit zwischen Lehrenden und Studierenden (als Moderator für prüfungsprozessinhärente Verzerrungen durch Stereotypisierung)
- das Zeitbudget für Lehre und deren Stellenwert (vs. Forschung)
- die Ausrichtung der Wissenschaft: rein vs. angewandt
- den Grad an Liberalität von Werten und Einstellungen

Diese Merkmale überschneiden sich teilweise und sind als kulturell verstandene Charakteristika zum Teil auf epistemologische Strukturen zurückzuführen. Unterschiede in der Erkenntnisproduktion zielen in Bezug auf den Prüfungsprozess zu einem Großteil auf die Vergleichbarkeit von Prüfungen ab, da sie den Grad der Standardisierung beeinflussen. Wissenschaftsorganisatorische Merkmale hingegen beziehen sich vor allem auf die Ausprägung und Stabilität von Selektionsneigungen.

3.2.2 Fachkulturen in der deutschen Hochschulforschung

Im Vergleich zur US-amerikanischen Forschung liegt der Beschäftigung mit Fachkulturen in Deutschland (mit einzelnen Ausnahmen, z.B. Multrus 2004; Windolf 1992) ein weitestgehend gemeinsames theoretisches Verständnis zugrunde. Die Existenz von Fachkulturen wird in der Regel als Folge der fachlichen Hochschulsozialisation verstanden, innerhalb welcher sich ein fachspezifischer Habitus im Sinne Bourdieus ausformt. Durch habituelles Handeln werden fachspezifische Kulturmuster immer wieder reproduziert (Huber 1991). Die Verwendung des Habituskonzepts in der Fachkulturforschung findet sich bei Bourdieu selbst in seiner Beschäftigung mit dem universitären Feld, das er anhand der Pariser Professor*innenschaft untersucht. Bourdieu verortet die naturwissenschaftliche, die philosophische, die medizinische und die juristische Fakultät entsprechend ihrer unterschiedlichen Ausstattung mit den verschiedenen Kapitalarten (Bourdieu 1983) im „sozialen Raum“ (ebd. 1985:9).
Er stellt bei der Auswertung verschiedener Indikatoren für kulturelles und ökonomisches bzw. soziales Kapital fest, dass zwei Hierarchielinien innerhalb des universitären Feldes existieren. Die eine dieser Linien verläuft entsprechend der Ausstattung mit ökonomischem Kapital von den Naturwissenschaften aufwärts über die philosophische und die juristische bis hin zur medizinischen Fakultät am oberen Ende, die andere in entgegengesetzter Richtung entsprechend der Ausstattung mit kulturellem Kapital. Die Indikatoren, die Bourdieu für die Messung der Kapitalsorten heranzieht, umfassen neben soziodemographischen Variablen unter anderem die Mitgliedschaft in Komitees, Auszeichnungen, die Anzahl an Publikationen und die Teilnahme an Konferenzen. In Jura wird dabei die höchste Teilnahmequote an Konferenzen erzielt, gefolgt von den Naturwissenschaften, Medizin und

schließlich von den Mitgliedern der Philosophischen Fakultät, welche andererseits die Publikationsrangliste anführen, in der Jura, Medizin und zuletzt die Naturwissenschaften folgen (Bourdieu 1988).

In Anlehnung an die Analyse innerakademischer habitueller Unterschiede Bourdieus bildet sich die deutsche Fachkulturforschung um Ludwig Huber aus. Huber zeigt in seinen Arbeiten zahlreiche Kriterien zur Bestimmung und Unterscheidung kultureller Muster von Fächern auf. Sie sind zu einem großen Teil auf studentische Fachkulturen (die in der US-amerikanischen Forschung keine große Relevanz besitzen) fokussiert, berücksichtigen teilweise aber auch strukturelle Unterschiede in Lehre und Forschung und die Rolle von Lehrenden. Die Charakterisierung der Unterschiede zwischen Fächern als kulturelle Unterschiede begründet Huber damit, dass eine rein inhaltliche bzw. epistemologische Bestimmung der Fächer die empirisch nachweisbaren Unterschiede zwischen den dort jeweils herrschenden „Wahrnehmungs-, Denk-, Bewertungs- und Handlungsmustern“ (Huber 1990a:72) nicht ausreichend erklären kann (ebd.) – eine Meinung, die nicht von allen Forscher*innen geteilt wird (zum Beispiel Arnold 2004).
Huber zeichnet in zahlreichen Schriften Unterschiede zwischen Fächern nach, wobei die von ihm verwendeten Differenzierungskriterien in ihrer Anzahl und in der Genauigkeit ihrer Ausführung variieren. Auch die empirischen Belege, mit denen er die Existenz der Unterschiede prüft, finden sich nur unregelmäßig eingestreut, meist in Form von Literaturverweisen, seltener in Form von eigenen Daten. Die Hauptkategorien, nach denen Huber Fächer unterscheidet sind die soziale Zusammensetzung der Studierenden, normative Klimata der Studierenden und Lehrenden, ihre Lebensstile und kulturelle Präferenzen, Interaktionsstrukturen innerhalb der Studierenden sowie zwischen Studierenden und Lehrenden, Lehrgestaltung und Lernsituationen, die zeitliche und räumliche Organisation des Lernens, der Curriculare Code und epistemologische Merkmale (Huber 1992).
Die fachspezifischen Unterschiede in der sozialen Zusammensetzung der Studierenden versteht Huber als Ergebnis der Selbstselektion der Studierenden bei der Fächerauswahl (Huber 1990b), als „Evidenz für fachspezifische Unterschiede der ‚Rekrutierung‘“ (ebd. 1991:440; auch ebd. 1990c). So entstammen Jurist*innen und Mediziner*innen besonders häufig Akademiker*innenfamilien, während sich in den Sozialwissenschaften und in der Psychologie in etwa ausgeglichene Verhältnisse zwischen Studierenden mit Akademiker*innenherkunft und Studierenden aus Arbeiter*innenfamilien zeigen (ebd. 1990b; schon Wilcke 1976: Untere Sozialschichten sind in Medizin und Jura unter-, obere Sozialschichten überrepräsentiert; aktuell Middendorfff et al. 2013: obere Schichten in Medizin, Musik, Psychologie am stärksten über-, in Betriebswirtschaftslehre (BWL), Bauingenieurswesen und Erziehungswissenschaften am stärksten unterrepräsentiert. In Jura ist der Anteil an Studierenden mit akademischem Bildungshintergrund rückläufig (Ramm et al. 2011)).

Die Kategorie der normativen Klimata dient als Oberkategorie für die (politischen) Einstellungen der Studierenden und Lehrenden, ihre Haltungen gegenüber der sozialen Ordnung und sozialen Strukturen und für ihr politisches und hochschulpolitisches Engagement (Huber 1992). Studentische Einstellungen umfassen dabei unter anderem Studien(fach)motive: In den Geistes- und Naturwissenschaften überwiegt das Fachinteresse, einhergehend mit dem Wunsch nach Bildung als Studienziel. In den Rechts- und Wirtschaftswissenschaften sollen Karriere, Einkommen, Macht und Prestige erreicht werden, das Studium stellt nur Mittel zum Zweck dar[13] (Liebau/Huber 1985; diese Einschätzung für Germanistik und Betriebswirtschaftslehre bestätigend: Gass/Meister: 1996).

In Hinsicht auf die politischen Einstellungen führen Huber sowie Huber und Liebau Befunde für eine fachspezifische Skala von links (Sozialwissenschaften) nach rechts (Rechtswissenschaften, Medizin, Ingenieurswissenschaften) für Lehrende und Studierende an (Huber 1990c; 1991; Liebau/Huber 1985), wie sie auch für US-amerikanische Fächer bekannt ist[14] (Gaff/Wilson 1971; Ladd/Lipset 1975). Zudem fassen Huber und Portele Belege dafür zusammen, dass auch innerhalb der Hochschullehrenden die „tiefgreifenden Unterschiede zwischen eher affirmativer und eher kritischer Einstellung zur sozialen Ordnung, in der man lebt“ (Huber 1990a:89) erwähnenswert sind, die mit der Verortung innerhalb der politischen Links-Rechts Skala korrespondieren (Huber/Portele 1983:213f).

Eine nicht unwesentliche Anzahl an Distinktionskriterien übernimmt Huber aus den Erkenntnissen der US-amerikanischen Forschung. So zum Beispiel auch die Grenzziehung zu anderen Fächern mittels der Bildung von Fach-Stereotypen, die Huber als Folge der fachlichen Sozialisation betrachtet. Sie fällt ebenfalls unter den Oberbegriff der normativen Klimata. Huber verweist zur empirischen Fundierung der Existenz fachspezifischer Stereotype jedoch lediglich auf Becher (Becher 1981; Huber 1990a:86f). Unterschiede in den Lebensstilen und kulturellen Präferenzen von Professor*innen sieht Huber durch Bourdieus Analyse der Pariser Fakultäten nachgewiesen (Bourdieu 1988:43ff; Huber 1990a:89). Hier sind zusätzlich zu den bereits erwähnten Einflussmöglichkeiten des Grads liberaler Werte und Einstellungen auf die Selektionsneigung kaum weitere systematische Einflüsse auf das Verhalten von Prüfenden abzuleiten.

Huber verweist zudem auf Befunde Apels und Englers, (siehe unten), die nach Fächern differenzierbare Vorlieben für bestimmte „‘hochkulturell‘-musische, populärmusische, […] sportliche, sowie reli-

[13] Zu unterschiedlichen Studienmotiven in den Fächern siehe bereits Wilke (1976): Studierende der Wirtschaftswissenschaften (auch Gleich et al. 1982) und in Jura sind selten intrinsisch motiviert (auch Ramm et al. 2011). Die höchste intrinsische Motivation findet sich in der katholischen Theologie, es folgen Geschichte und Germanistik.

[14] Siehe auch Maier-Leibnitz/Schneider (1991): Soziolog*innen und Politikwissenschaftler*innen stufen sich als politisch links ein, über Pädagog*innen, Philosoph*innen, Anglist*innen und Architekt*innen geht es auf der Einstellungsskala hin zu den sich eher als rechts verstehenden Ingenieurswissenschaftler*innen, Mediziner*innen, Jurist*innen und Agrarwissenschaftler*innen. In der US-amerikanischen Skala finden sich kleinere Abweichungen von den deutschen Verhältnissen, u.a. bedingt durch einen unterschiedlichen Status der Rechtswissenschaften.

giöse und soziale Aktivitäten […], Ernährungsweisen, Kleidungsstile […], Wohnformen“ (Huber 1991:440) bei Studierenden aufzeigen. Diesen „‘Äußerlichkeiten‘ der Lebens- und Arbeitsformen“ (Liebau/Huber 1985:316), wie etwa Kleidungsstil, Wohnform, und Einrichtung bzw. Gestaltung universitärer Räumlichkeiten (Huber 1990c) kommt laut Huber Bedeutung zu, da sie als Ausdruck bestimmter Haltungen interpretiert werden können (ebd.). Für den Notengebungsprozess dürften derartige äußerliche Kriterien keine eigene Rolle spielen, sie sollten sich weitestgehend durch soziodemographische Merkmale der Zusammensetzung von Studierenden sowie durch fachspezifische Rekrutierungsmuster erklären lassen (dazu auch Apel 1989).

Für die Analyse von Interaktionsstrukturen führt Huber die Kriterien Kontaktdichte, Art der Kontakte, Kontaktrichtung[15], Betreuungsrelation, Sprachform, Grad der Arbeitsteilung und Kommunikationsreichweite an (Huber,1990a; 1992:102; Liebau/Huber 1985). Er verweist auf eine Untersuchung von Gleich et al., die 1982 den häufigsten Kontakt zwischen Studierenden und Lehrenden in den Sprachwissenschaften feststellen, während die Wirtschaftswissenschaften den geringsten Kontakt zwischen den beiden Gruppen aufweisen (Gleich et al. 1982). Hinsichtlich der Betreuungsrelation errechnet Huber die höchste Studierende-pro-Lehrende Quote für die Fächergruppe Sozial- und Wirtschaftswissenschaften/Jura, gefolgt von Kunst und den Sprachwissenschaften. In der Medizin kommen die wenigsten Studierenden auf einen Dozierenden (Huber 1990b)[16]. Hier ist ein Einfluss auf die Lernbedingungen zu erwarten, womit bei besseren Betreuungsrelationen auch bessere Prüfungsleistungen zu erwarten sind.

Die Analyse der Arbeitsteilung widmet sich der Interaktion innerhalb der scientific community. Sie kann sowohl auf einer horizontalen als auch auf einer vertikalen Ebene erfolgen. Die horizontale Ebene der Arbeitsteilung umfasst die beiden Gegensätze a) starke Arbeitsteilung mit wenig gegenseitigem Erklärungsbedarf hinsichtlich der wissenschaftlichen Grundlagen, dem Vorgehen und den Methoden versus b) niedrige Arbeitsteilung mit viel gegenseitigem Erklärungsbedarf. Auf der vertikalen Ebene kann die Unterscheidung hierarchische Struktur versus egalitäre Teamarbeit getroffen werden (Huber 1990a). Es wird schnell deutlich, dass Huber sich bei der Beschreibung der horizontalen Ebene

[15] Siehe zur Kontaktrichtung auch Apenburg et al. (1977b): In Betriebswirtschaftslehre ist nicht nur der Kontakt zu den Lehrenden am seltensten und als schlecht bewertet, auch der Kontakt zwischen den Studierenden wird im Vergleich zu anderen Fächern als lockerer wahrgenommen. In Medizin, Mathematik und Physik und besonders in Chemie wird der Kontakt zu anderen Studierenden als enger empfunden.
Dass sich dieses Muster im Laufe der Zeit nur gering gewandelt hat, zeigt sich bei Multrus et al. (2008): Kontakte zwischen Studierenden des eigenen Fachs sind an Universitäten in Medizin und den Ingenieurswissenschaften am häufigsten, es folgen die Natur- vor den Wirtschafts- und Rechtswissenschaften, der geringste Kontakt findet sich in den Kultur- und Sozialwissenschaften

[16] Ein Überblick über die langfristige Entwicklung der Betreuungsrelationen nach Fächergruppen findet sich in Lundgreen et al. (2009). Dieser bestätigt die von Huber errechnete Rangfolge und zeigt, dass sie auch langfristig stabil ist. Die von Lundgreen et al. präsentierten Daten verdeutlichen zudem, dass sich die Anteile des Beschäftigungsstatus von Lehrpersonal (Professor*innen vs. akademische Mitarbeiter*innen vs. Lehrkräfte für besondere Aufgaben) sich zwischen den Fächergruppen unterscheiden.

der Arbeitsteilung auf Whitley bezieht, denn im Grunde beschreibt er hier die im Falle a) hohe und im Falle b) niedrige funktionale Abhängigkeit zwischen Forschenden (Whitley 1984).

Empirische Nachweise für deutsche Verhältnisse bleibt Huber an dieser Stelle ebenso schuldig wie für die postulierten Unterschiede in der Kommunikationsreichweite zwischen den Fächern und im jeweiligen Wissenschaftsverständnis. Stattdessen orientiert er sich erneut an der US-amerikanischen Forschung und verweist in Bezug auf die Reichweite der Netzwerke explizit auf die von Gouldner getroffene Unterscheidung zwischen Cosmopolitans und Locals (Gouldner 1957; Huber 1990a) und in Hinblick auf das Wissenschaftsverständnis auf die Unterscheidungen nach reinen versus angewandten und harten versus weichen Fächern. Diese Unterscheidungen, sowie die Kombination dieser Kriterien wie bei Biglan und Becher (Becher 1987b; Biglan 1973) stellen seiner Meinung nach ein nützliches Analyseinstrument dar, um Muster der Problemstellung und -bearbeitung, Gütekriterien wissenschaftlicher Arbeit und die jeweils vorherrschenden Methoden des Zugriffs auf die Wirklichkeit zu erfassen (Liebau/Huber 1985).

Unterschiede in der Häufigkeit der Kontakte zwischen Studierenden und Lehrenden, die sich entsprechend der fachlichen Zugehörigkeit zeigen (zusammenfassend Huber 1991) sollten nach Huber immer auch hinsichtlich der Qualität der Kontakte bewertet werden, ob sie also innerhalb oder außerhalb der Lehrveranstaltungen stattfinden, ob es sich um akademische oder private Themen handelt (Liebau/Huber 1985). Auch ob die Studierenden vorwiegend Kontakt untereinander oder zu den Lehrenden pflegen und ob sie sich in Fachsprache oder umgangssprachlich unterhalten variiert nach Huber von Fach zu Fach (Huber 1990b 1991), ebenso wie die quantitative Betreuungsrelation (siehe oben).Dadurch werden unterschiedliche Interaktionsstrukturen in den einzelnen Fachkulturen gefördert. Diese Beeinflussung der Interaktion zwischen Studierenden und Lehrenden führt zur nächsten Differenzierungslinie Hubers, der Lehrgestaltung und den Lernsituationen.

Hier verweist Huber auf Befunde, die fachspezifische Verhältnisse von Veranstaltungsformen[17] (Vorlesungen, Übungen, Seminaren, Tutorien und Praktika) und Unterschiede in den Diskussionsstilen (offen/geschlossen; auf Konvergenz oder Divergenz ausgelegt; Bluff begünstigend/ erschwerend, dazu auch Reiss 1975), in den Ausrichtungen der Lehrveranstaltungen (stoffzentriert/ personenzentriert; leistungszentriert/interessenzentriert), in den Lehrstilen[18] (strukturiert/ flexibel; mediengestützt/frei), in den Partizipationschancen[19] (hoch/niedrig) sowie in den Lernanforderungen (hierar-

[17] Siehe unter anderem Oehler et al. (1976): Der Anteil von Vorlesungen an allen Veranstaltungen ist in Maschinenbau besonders hoch, es folgen die Wirtschaftswissenschaften vor Architektur und dem Lehramtstudium. Lehre findet in Letzterem vor allem in Tutorien und Seminaren statt, was in Architektur und Maschinenbau deutlich seltener der Fall ist.

[18] Siehe auch Portele (1976): Während der Lehrstil in Mathematik, Archäologie und Elektrotechnik durch Monologe der Dozierenden geprägt ist, dominieren in Soziologie und in den Politikwissenschaften Diskussionen die Kommunikation im Lehrbetrieb.

[19] Siehe auch Weigand (2012): In den Kulturwissenschaften herrscht eine hohe Autonomie der einzelnen Forschenden, dem steht ihre Abhängigkeit in den Naturwissenschaften gegenüber. Die dort herrschenden hierar-

chisch/zyklisch; logisch/interpretierend; fremd/selbstbestimmt; regelgeleitet/ diffus) und den daran angepassten Lernstrategien ausarbeiten (Huber 1991; auch: Kolb 1981). Er betont auch den von Fach zu Fach wechselnden Stellenwert von Lehre im Vergleich zu Forschung[20] und Praxis und unterschiedliche Aufgaben von Lehre (Unterstützung der Wissensinternalisierung/Unterstützung von Problemlösungskompetenzen), wiederum ohne konkrete Belege für fachspezifische Unterschiede vorzulegen (Huber 1992). Auch die Überprüfung des Gelehrten findet Beachtung bei Huber. Er erkennt in den von ihm hauptsächlich in ihrer Form als Selektionsinstrumente im Studium betrachteten Prüfungen unterschiedliche Prüfungsformen - „'Fakten-Pauken'" und „'Oberflächen-Lernen'" gegenüber „'Tiefen-Lernen' oder strukturellem Verstehen" (Huber 1991:423f) – die es auf Unterschiede zwischen den Fächern zu prüfen gilt. Er verweist für den Nachweis fachspezifischer Unterschiede in Prüfungen an dieser Stelle auf Teichler (1987)[21].

---

chischen Strukturen, die sich etwa aus der Notwendigkeit der Nutzung technischer Geräte ergeben sowie der kumulative Charakter der Wissensbildung lassen dem Individuum ein vergleichbar geringes Maß an Entscheidungsfreiheit im Forschungsbetrieb zu. Entsprechend ist auch der Lehrbetrieb in den Naturwissenschaften deutlich hierarchischer strukturiert.

[20]Auch Schimank (1992): In den Geistes- und Sozialwissenschaften ist der zeitliche Aufwand für Forschung am geringsten, gefolgt von den Agrarwissenschaften und Medizin. In den Natur- und Ingenieurswissenschaften ist der höchste zeitliche Forschungsaufwand zu verzeichnen. Vor allem in der Medizin polarisiert sich die Forschungslage in Viel- und Wenigforschende (über 40% bzw. unter 10% Anteil Forschung am Zeitbudget), in den Ingenieurswissenschaften zeigt sich der größte Anteil Vielforscher*innen.

Von Kopp und Weiß (1995) präsentieren eine ähnliche Rangfolge: Auch hier weisen die Ingenieurswissenschaften den höchsten Zeitaufwand für Forschung (am gesamten Zeitbudget) auf, gefolgt von den Naturwissenschaften und Mathematik. Allerdings findet sich hier der geringste Forschungsaufwand in der Gruppe der Sprach- und Kulturwissenschaften, während in den Sozialwissenschaften mit Rechts-und Wirtschaftswissenschaften zusammengefasst noch etwas mehr Zeit aufgewendet wird. Mit sinkendem Anteil der Zeit, die auf Forschung verwendet wird, steigt in der gleichen Rangfolge der Anteil des Zeitbudgets, der für Lehre genutzt wird.

Enders und Teichler (1995) ermitteln damit übereinstimmend, dass nur in Pädagogik und den Kunstwissenschaften mehr Zeit für Lehre als für Forschung verwendet wird, in den Sprach-, Kunst-, Wirtschafts-, Natur- und besonders in den Rechtswissenschaften verhält es sich umgekehrt und in den Sozial- und Ingenieurswissenschaften halten sich die Anteile die Waage. Erziehungs- und Kunstwissenschaftler*innen investieren auch mehr Zeit in Lehre, als es in anderen Fächern der Fall ist und halten die meisten Lehrveranstaltungen ab. In den Naturwissenschaften und in Medizin finden die wenigsten Lehrveranstaltungen statt. Entsprechend wird in den Erziehungs- und Kunst- aber auch in den Wirtschaftswissenschaften die Präferenz eher auf Lehre als auf Forschung gesetzt, während in den übrigen Fächern vornehmlich der Forschung persönlicher Vorrang vor der Lehre eingeräumt wird.

Jacob und Teichler (2011) bestätigen zudem, dass Vertreter*innen der Geistes und Sozialwissenschaften mehr Zeit für Lehre und weniger Zeit für Forschung aufbringen als die der Natur- und Ingenieurswissenschaften. Entsprechend ist auch die Drittmittelantragsaktivität in den Geisteswissenschaften am geringsten, in der Medizin und den Naturwissenschaften am höchsten (Böhmer et al. 2011).

[21] Dippelhofer-Stiem (1983) charakterisiert die Wirtschaftswissenschaften und Medizin als ausgerichtet auf den Erwerb von Faktenwissen, natur- und vor allem ingenieurswissenschaftliche Fächer als stärker verständnisorientiert. In Germanistik steht hingegen das Verfahren des wissenschaftlichen Arbeitens im Vordergrund.

Schulz et al. (2014) identifizieren eher anwendungsorientiert prüfende (naturwissenschaftliche und technische Studiengänge), eher wiedergabeorientiert prüfende (Medizin und Sportwissenschaften) und gemischt prüfende (sozialwissenschaftliche Studiengänge, Architektur sowie Ernährung, Landnutzung und Umwelt).

Hinsichtlich der Anforderungen in (Zwischen-)Prüfungen zeigen Apenburg et al. (1977b), dass sie in Jura am höchsten wahrgenommen werden. Es folgen Mathematik/Informatik, Chemie und Physik. In Germanistik, Betriebswirtschaftslehre, Französisch und Medizin werden die Anforderungen in zunehmendem Maße als geringer eingestuft.

Als eng verknüpft mit der Lehrgestaltung und den Lernsituationen und Lernstilen sieht Huber die jeweilige zeitliche und räumliche Arbeitsorganisation in den Fächern. Hier stellt er Fächer, in denen häufige räumliche Präsenz, beispielsweise in Laboren, und zeitliche Pflichtveranstaltungen zusammenfallen und die damit eine Anpassung der studentischen Selbstorganisation an die akademischen Strukturen fordern solchen Fächern gegenüber, in denen die Studierenden Ort und Zeit des Lernens relativ frei bestimmen können (Huber 1991).

Curriculare Codes tragen in ihrer jeweiligen Ausprägung ebenfalls zu unterschiedlich ausgestalteten Fachkulturen bei. Dieses Differenzierungskriterium geht auf Basil Bernstein zurück, der zwischen zwei Typen von Curricula unterscheidet, zwischen dem „Sammlungstyp“ (Bernstein 1977:128), der sich durch disziplinär klar abgegrenzte Inhalte auszeichnet, und dem „integrierten Typ“ (ebd.), für den eine Offenheit der Inhalte gegenüber anderen Betrachtungsweisen charakteristisch ist. Formen der deutlichen Abgrenzung von Inhalten bezeichnet Bernstein auch als strenge Klassifikation, schwache Grenzen zwischen Inhalten bedeuten schwache Klassifikation (ebd.). In Kombination mit der einem Fach zuzuordnenden „Rahmung“ (ebd.), die den Spielraum im Prozess der Vermittlung von Inhalten beschreibt und nach Bernstein ebenfalls entweder schwach (bei geringem Gestaltungsspielraum) oder stark (bei hohem Gestaltungsspielraum) ausfallen kann, ergibt sich für die jeweilige Ausprägung von Klassifikation und Rahmung ein entsprechender curricularer Code.

Curriculare Codes können als Sammlungs- bzw. Kollektionscode oder als Integrationscode auftreten. Kollektionscodes sind durch starke Klassifikation und strenge Rahmung geprägt (ebd.), also durch „scharfe Differenzierung und hierarchische Strukturierung der Inhalte nach innen und starke Abgrenzung derselben nach außen“ und sie lassen „wenig Raum für Mitgestaltung des Curriculums“ (Huber 1991:438). Integrationscodes sind vor allem durch schwache Klassifikation gekennzeichnet, während die Stärke der Rahmung variieren kann (Bernstein 1977). Sie weisen also weniger stark abgegrenzte Inhalte auf und gestatten grundsätzlich einen höheren Gestaltungsspielraum in der Lehre. Huber verweist zwar auf Unterschiede zwischen Fächern hinsichtlich des dort herrschenden curricularen Codes, belegt sie aber ebenfalls nicht weiter (Huber 1991). Weigand (2012) ordnet den Naturwissenschaften den Kollektionscode und damit einhergehend die Vermittlung hierarchisch strukturierter und kumulativer Wissensbestände, den Kulturwissenschaften und der dortigen korrigierend verlaufenden und eher interpretationsbasierten Wissensvermittlung den Integrationscode zu.

Huber hebt wie auch Becher (1989) unterschiedliche Beziehungen der Fächer zur gesellschaftlichen Umwelt des Hochschulsystems hervor. So weist Huber auf die Relevanz der Beziehungen von Jurist*innen, Ingenieur*innen und Mediziner*innen zu den entsprechenden Professionen dieser Fächer hin und stellt dem dort (und auch in den Wirtschaftswissenschaften) herrschenden Berufsbezug den Wissenschaftsbezug in geistes- und naturwissenschaftlichen Fächern entgegen (Liebau/Huber 1985). Explizit verweist er auf die fächerspezifischen Auswirkungen von Arbeitsmarktentwicklungen (Huber

1991; zu fächerspezifischen Arbeitsmarktchancen im Einzelnen siehe zum Beispiel: Grotheer et al 2012; Parmentier et al. 1998; Reisz/Stock 2013; Stief/Abele 2002; Wissenschaftsrat 1999).
In erneut starker Anlehnung an Bourdieu sieht Huber schließlich die Bedeutung des jeweiligen Fachs in der Gesellschaft als relevantes Differenzierungskriterium. Die Verortung von Fächern im sozialen Raum anhand von Variablen des sozialen Status und Indikatoren wie etwa der Einbindung eines Faches in die Praxis einer Profession (Huber 1990a) oder der Höhe von eingeworbenen Drittmitteln[22] (Huber 1990b) ist seiner Meinung nach geeignet, die gesellschaftliche Relevanz der einzelnen Fächer genau zu bestimmen.

Es wird schnell deutlich, dass Huber sich in der Ausarbeitung von Fachunterschieden stark an den Merkmalen der US-amerikanischen Forschung orientiert. Und auch neuere Unterscheidungsansätze, wie die Differenzierung nach curricularem Code oder die Partizipationschancen weisen wieder Zusammenhänge mit bereits bekannten Distinktionen auf (hier starke Abgrenzung des kumulativen Wissens in harten Fächern mit hohem Standardisierungsgrad in Lehr- und Lernumgebungen vs. schwache Abgrenzung des korrigierenden Wissens in weichen Fächern mit geringem Standarisierungsgrad). Hubers Arbeiten bieten dennoch einen neuen Ansatzpunkt, der vor allem im Hinblick auf den Notengebungsprozess eine Rolle spielen dürfte: Die Zusammensetzung der Studierenden in Bezug auf *leistungskonforme* Ursachen. Wird die bei Huber angesprochene soziale Herkunft der Studierenden berücksichtigt, besteht die Möglichkeit, dass etwa Akademiker*innenkinder durch die Kenntnisse, die ihre Eltern vom Hochschulsystem besitzen, Vorteile in ihren Lernstrategien, in ihrer Kurswahl usw. erhalten. Auch fachspezifische Lernmotivationen könnten durch unterschiedliche Lernstrategien unterschiedliche Noten begünstigen: So könnte eine intrinsische Bildungsmotivation einerseits weniger taktische Entscheidungen für oder gegen bestimmte Lern- und Kurswahlstrategien beinhal-

[22] Von Kopp und Weiß (1995) finden Unterschiede in der Bedeutung verschiedener Drittmittelquellen für verschiedene Fächergruppen: Die wichtigste Drittmittelgeberin für alle Gruppen stellt die DFG dar. In den Ingenieurswissenschaften kommen Bund und Ländern sowie der Privatwirtschaft, in der Gruppe der Rechts-, Wirtschafts- und Sozialwissenschaften Stiftungen eine größere Bedeutung zu als in den anderen Gruppen. Für die Ingenieurswissenschaften stellen Drittmittel außerdem relativ betrachtet einen größeren Teil der Gesamtfinanzierung als für die restlichen Fächergruppen dar.
Enders und Teichler (1995) präsentieren den Befund, dass die Naturwissenschaften hinsichtlich der erhaltenen Forschungsgelder vor den Ingenieurswissenschaften an der Spitze aller Fächergruppen liegen. Rechts- und Wirtschaftswissenschaften belegen hier die letzten Plätze. Erziehungswissenschaftler*innen erhalten ihre Forschungsgelder dieser Studie nach vor allem von staatlichen Institutionen und den Hochschulen, wohingegen Mediziner*innen und Wirtschaftswissenschaftler*innen ihre Forschungsgelder am häufigsten aus der Privatwirtschaft generieren. Die übrigen Fächergruppen beziehen Drittmittel meist aus öffentlichen Einrichtungen der Forschungsförderung.
Jacob und Teichler (2011) ergänzen diese Befunde um die Erkenntnis, dass die eher praxisorientiert und kommerziell ausgerichtet forschenden Natur- und Ingenieurswissenschaftler*innen häufiger Drittmittel aus der öffentlichen Forschungsförderung und von Privatunternehmen erhalten als die gesellschaftsorientiert forschenden Geistes- und Sozialwissenschaftler*innen, die dafür häufiger Förderung von der eigenen Institution erhalten.

ten als ein rein extrinsischer, auf die Verwertung des Abschlusses ausgelegter Antrieb. Andererseits ist es denkbar, dass mit größerem Interesse an den Studieninhalten als am Abschluss auch geringere Mühen bei der Wissensaufnahme und damit bessere Lernleistungen einhergehen.

Hubers Verortung von Fachkulturen im sozialen Raum stellt den Anknüpfungspunkt für eine Reihe von Untersuchungen zur fachkulturellen Habitusausbildung dar, die sich auf die Vorarbeit der Forscher*innen um Huber beziehen (vgl. etwa Apel (1989) zu studentischen Lebensstilen, Frank (1990) zum Ausdruck eines spezifischen Habitus im fachlichem Selbstverständnis, Engler (1993) zur Reproduktion geschlechtsspezifischer Ungleichheiten in Fachkulturen und Schaeper (1997) zu geschlechts- und fachspezifischen Unterschieden in Lehrkulturen).
Relevanz für den Benotungsprozess besitzen diese Anschlussarbeiten nur begrenzt, wenn etwa epistemologische Merkmale oder die Ausstattung mit finanziellen Mitteln und damit einhergehend die Abhängigkeit von der gesellschaftlichen Umwelt (ebd.) thematisiert wird. So beschreibt Engler die Erziehungswissenschaften als junges, mit geringen finanziellen Mitteln und ohne einheitliches Grundlagenwissen ausgestattetes Fach, dessen Fachkultur vor allem auf kulturellem Kapital beruht. Die Rechtswissenschaften werden von ihr als traditionelles Fach, mit ausgeprägtem Praxisbezug, einem einheitlichen Wissensfundus und ebenfalls begrenzten (akademischen) finanziellen Mitteln beschrieben, in dem sich kulturelles und ökonomisches Kapital die Waage halten. Die beiden Ingenieurswissenschaften Maschinenbau und Elektrotechnik sind vom stärksten Praxisbezug, einem enormen finanziellen Hintergrund und einer internationalen Ausrichtung gekennzeichnet, was ein deutliches Mehr an ökonomischem Kapital im Vergleich zu kulturellem Kapital nach sich zieht[23] (Engler 1993).
Auch die Lern- und Lehrgestaltung unterscheidet sich zwischen den untersuchten Fächern. Studierende der Elektrotechnik gefolgt von denen des Maschinenbaus verbringen mehr Zeit in Lehrveranstaltungen als die in den nahe beieinander liegenden Fächern Pädagogik und Jura (bei Oehler et al. (1976) finden sich Maschinenbau, (Sozial-)Pädagogik und Jura in der gleichen Reihenfolge). Maschinenbauer*innen sind im Hinblick auf das Selbststudium und die insgesamt in das Studium investierte Zeit am stärksten ausgelastet, die Jurist*innen liegen beim Selbststudium auf dem zweiten und insgesamt auf dem dritten Platz, jeweils im Wechsel mit den Elektrotechniker*innen. Die Erziehungs-

[23] Auch Schölling (2005) verortet verschiedene Fächer anhand der jeweiligen Kapitalausstattung im sozialen Raum: Jura, Medizin, Chemie, Mathematik, Physik, Elektrotechnik und Maschinenbau befinden sich mit in dieser Reihenfolge abnehmendem Gesamtkapitalvolumen auf der Seite der Fächer mit hohem ökonomischem und geringem kulturellen Kapital. Studiengänge mit Abschluss Magister, Lehramtsstudiengänge, Pädagogik und sozialwissenschaftliche Fächer befinden sich mit in dieser Reihenfolge abnehmendem Gesamtkapitalvolumen auf der Seite der Fächer mit niedrigem ökonomischem und hohem kulturellem Kapital. Die Wirtschaftswissenschaften und Psychologie werden sowohl hinsichtlich des Gesamtvolumens als auch hinsichtlich der Kapitalsorten im mittleren Bereich des Raums verortet.

wissenschaften liegen hier jeweils an letzter Stelle (bei Oehler et al. (1976) liegen die Jurist*innen im Selbststudium vorne[24]).

Die fachspezifische Rekrutierung sieht Engler unter anderem durch die Schulkarriere begründet. So haben Studierende der beiden ingenieurswissenschaftlichen Fächer überdurchschnittlich häufig Kurse in Mathematik, Werken und den Naturwissenschaften belegt, in denen erziehungswissenschaftliche Studierende unterdurchschnittlich häufig zu finden waren. Letztere haben wiederum in der Schule deutlich häufiger Pädagogik, Deutsch, Religion und Gesellschaftslehre gewählt, wo die späteren Ingenieurswissenschaftler*innen unterrepräsentiert waren. Die Studierenden der Rechtswissenschaften fanden sich zu Schulzeiten ebenfalls besonders häufig in den Fächern Deutsch und Gesellschaftslehre sowie in Geschichte – in der Vielzahl der Fächer, in denen Erziehungs- oder Ingenieurswissenschaftler*innen deutlich über- bzw. unterrepräsentiert waren, lagen sie anteilsmäßig zwischen diesen Gruppen (auch: Schölling 2005; Ramm et al. 2011).

Schaeper (1997) nimmt Lehrveranstaltungen als einen zentralen Aspekt der Lehre in den Fächern Germanistik, Pädagogik, Wirtschaftswissenschaften, Biologie und Physik in den Fokus. Schaepers Analysen hinsichtlich fachspezifischer Unterschiede zeigen, dass in Pädagogik, Germanistik und vor allem in den Wirtschaftswissenschaften mehr Zeit auf Lehre verwendet wird als in den beiden naturwissenschaftlichen Fächern[25] (ebd.; ebd. 1995). Der Stellenwert von Lehre im Verhältnis zur Forschung ist bei Lehrenden der Pädagogik am höchsten, gefolgt von Germanistik und den Wirtschaftswissenschaften. In Biologie und Physik wird der Lehre im Vergleich zur Forschung der geringste Wert beigemessen (ebd. 1995). Germanist*innen und Pädagog*innen berichten zudem von der größten Gestaltungsfreiheit in den Lehrveranstaltungen (ebd.). Auch hinsichtlich der Lernziele und Lehrpraktiken gibt es deutliche Unterschiede zwischen den Fächern. Lehrende der Pädagogik legen hohen

[24] Siehe aktuell auch Ramm et al. (2011) für eine Übersicht über den Zeitaufwand für Lehrveranstaltungen und Selbststudium in 25 Fächern. Hier liegen Studierende der Pharmazie bei Lehrveranstaltungen an der Spitze, gefolgt von Chemie- und Medizinstudierenden, in den Politikwissenschaften wird hierfür die wenigste Zeit verwendet. In Medizin wird die meiste Zeit für das Selbststudium (gefolgt von Physik) und insgesamt für das Studium (etwa gleichauf liegt die Pharmazie) aufgebracht. Im Widerspruch zu den Befunden Englers berichten Studierende im Maschinenbau hier von dem geringsten Zeitaufwand für das Selbststudium, in den Politikwissenschaften wird insgesamt am wenigsten Zeit für das Studium aufgebracht.
Bei Gawatz (1991) ist der Zeitaufwand für Lehrveranstaltungen in der Medizin am höchsten, es folgen die Natur- und Ingenieurs-, dann die Wirtschaftswissenschaften, Kulturwissenschaften und schließlich die Rechts- und Sozialwissenschaften. Im Selbststudium liegen hier Jurastudierende vorne, die der Sozialwissenschaften am Ende. Insgesamt berichten Studierende der Medizin vor denen der Natur-, dann der Rechts- und dann der Ingenieurswissenschaften von dem höchsten Zeitaufwand fürs Studium. Die wirtschafts-, kultur- und sozialwissenschaftlichen Studierenden bilden den Abschluss der Rangfolge.

[25] Ipsen (1976) ermittelt in einer Zeitbudgetanalyse der wissenschaftlichen Beschäftigten in 11 Fächern unter anderem unterschiedlich hohe Arbeitszeiten während des Semesters (am geringsten in Jura, am höchsten in Medizin), Unterschiede im Aufwand für Forschung (am geringsten in Anglistik, am höchsten in Medizin), Lehre (am geringsten in Medizin, am höchsten in Anglistik) und Verwaltung (nur geringe Unterschiede, höchster Aufwand in Soziologie), wobei sich ein grundsätzliches Muster zeigt, dass ein vergleichsweise geringes Maß an Forschung mit mehr Lehrzeit einhergeht und umgekehrt.

Wert auf Praxisbezug, Kritikfähigkeit und soziale Kompetenzen und sind am ehesten bemüht, die Studierenden in den Mittelpunkt der Veranstaltung zu stellen. Germanist*innen heben ebenfalls die beiden letzten Lernziele hervor und stehen bei den Bemühungen ihre Veranstaltungen auf die Studierenden zu zentrieren an zweiter Stelle, messen dem Praxisbezug allerdings einen geringen Wert bei. Der spielt auch in der Volkswirtschaftslehre (VWL) keine große Rolle, genauso wie die sozialen Kompetenzen, denen in diesem Fach die geringste Bedeutung zukommt. Zudem sind wirtschaftswissenschaftliche Lehrende am wenigsten bemüht, die Lernenden in den Mittelpunkt der Veranstaltungen zu stellen, in Physik und Biologie herrscht diesbezüglich ähnliches Desinteresse (ebd.).

Wie auch bei Huber selbst weisen die an seine Arbeiten anschließenden Studien eine solide theoretische Fundierung auf. Der Fokus liegt dabei häufig auf Aspekten gesondert gefasster studentischer Fachkulturen oder auf der Lehre. Merkmale, aus denen sich Einflüsse auf den Benotungsprozess ableiten ließen, sind nur solche behandelt, die auch bereits im Kontext der US-amerikanischen Forschung besprochen wurden. Gleiches lässt sich auch für die Arbeit von Köhler und Gapski (1997) sagen, die im Rahmen des Forschungsprojekts „Lebenswelt Studierender" die drei Fächer Chemie, Biochemie und Geschichte anhand der vier Kategorien Diskursformation, Handlungsfeld im sozialen Raum, typischer Habitus in Forschungs- und Lehrkultur und typischer Habitus in der Lernkultur der Studierenden differenzieren (Köhler/Gapski 1997). Unter die Kategorie Diskursformation fassen sie die Unterscheidungen nach harten versus weichen Fächern, nach vorwiegender Orientierung auf Forschung im Gegensatz zur Anwendungsorientierung, nach dem jeweiligen Maß an konsentiertem Wissen, nach der Anzahl leitender Paradigmen und deren Stabilität, nach Dichte und Umfang von Kommunikationsnetzwerken, nach vorherrschenden Aufbereitungsformen von Forschungsergebnissen (häufige Publikationen versus in die „'Tiefe'" (ebd.:208) gehende Publikationen), sowie nach regionalen, nationalen oder internationalen Kommunikationsräumen. Diese Liste liest sich wie eine Zusammenfassung der fachlichen Unterscheidungskriterien aus der US-Literatur.

Alternativ zur Erfassung von Fachkulturen als Ausdruck habitueller Muster der Wahrnehmung, des Denkens und des Handelns, begreift Windolf sie als „System von Normen und Werten […] die bei der selektiven Produktion, Verteilung und Aneignung von Wissen zu beachten sind" (Windolf 1992:77). Er untersucht die These, dass Studierende ein bestimmtes Fach gewählt haben, weil ein Zusammenhang zwischen ihren „kognitiven Orientierungen und der Fachkultur" (ebd.:76) besteht. Der sozialisatorische Einfluss von Fachkulturen auf Studierende wird von ihm als vergleichsweise gering eingeschätzt. Er geht davon aus, dass Studierende sich vielmehr schon vor Aufnahme des Studiums in ihren jeweiligen Einstellungen und Dispositionen ähneln bzw. unterscheiden, was dann zu einer bestimmten Fachwahl führt.

Um diese Annahme zu prüfen, differenziert Windolf Fachkulturen hinsichtlich der „praktischen Verwertbarkeit des Wissens, dem Grad der Spezialisierung und der vorherrschenden Handlungsrationalität“ (ebd.). Die Ausprägungen aller drei Kriterien lassen sich in Windolfs Sinne kategorial erfassen, als wissenschaftsorientiert versus anwendungsorientiert, als in hohen Maße versus in niedrigem Maße spezialisiert sowie als wertrational versus zweckrational motiviert, wobei sie empirisch durchaus auch graduell bestimmbar sein sollten. Windolf versteht jede Fachkultur als Ergebnis eines Zusammenspiels dieser drei Kriterien, wobei sich durch die verschiedenen Ausprägungen bei kategorialer Auffassung insgesamt acht unterschiedliche Idealtypen bilden ließen. Windolf hingegen generiert in Anlehnung an Parsons und Platt (1973) ein Vier-Felder-Schema, in dem er Praxis- und Theorie als Ausrichtungen wissenschaftlicher Tätigkeit gegenüberstellt und diesen beiden Ausprägungen jeweils einen hohen und einen niedrigen Grad an Spezialisierung zuordnet. Die vorherrschende Handlungsrationalität wird in diesem Schema nicht als eigenständiges Differenzierungskriterium, sondern als mit den „Typen kognitiver Rationalität“ (Windolf 1992:76) verbunden betrachtet.

Um die These der weitgehenden Übereinstimmung zwischen den Werten eines Faches und den Einstellungen der Studierenden dieses Fachs noch vor Beginn der fachlichen Sozialisation zu überprüfen, nutzt Windolf die Antworten von Studienanfänger*innen der Universitäten Heidelberg und Saarbrücken sowie der TH Karlsruhe im Wintersemester 1989/90 auf Fragen zur Studienmotivation, gegliedert in die Komplexe Karriereorientierung, Studium als Lebensstil, (Selbst-)Reform und Wissenschaftsorientierung. Er erstellt daraus fachspezifische Antwortprofile, die die Zustimmung bzw. Ablehnung zur Grundausrichtung der Aussagen der einzelnen Themenkomplexe abbilden. Die Unterschiede zwischen den Fächern und damit die Zuordnung zu unterschiedlichen Fachkulturen erfolgt also über die Einstellungsmuster der Studienanfänger*innen.

Das faktorenanalytische Vorgehen zielt dabei auf die Hervorhebung von Unterschieden im Grad der Zustimmung bzw. Ablehnung zwischen den Fächern, die sich auf jedem Faktor in unterschiedlichem Ausmaß zeigen, so dass zwar Ähnlichkeiten zwischen Fächern in einzelnen Einstellungskomplexen auftreten, sich aber jede Fächergruppe in der Kombination aller vier Komplexe von den anderen unterscheidet. Es zeigt sich, dass Betriebswirtschaftler*innen und Maschinenbauer*innen stark positiv auf dem Karrierefaktor laden, auf dem Politolog*innen, Soziolog*innen, Pädagog*innen und Theolog*innen die höchsten negativen Werte aufweisen. Die Sozialwissenschaftlichen Fächer weisen dafür deutlich positive Werte in der Kategorie Studium als Lebensstil auf, hier liegen naturwissenschaftliche und Ingenieursfächer im negativen Bereich. Hohe Reformwerte erreichen Studierende der Theologie, der Psychologie, Politik und Pädagogik, am anderen Ende finden sich die der Elektrotechnik und Mathematik. Wissenschaftsorientierung findet sich vorwiegend in den naturwissenschaftlichen Fächern, am wenigsten in Volks- und Betriebswirtschaftslehre. Benachbarte Fächer wie Volkswirtschaftslehre und Betriebswirtschaftslehre oder auch Politologie und Soziologie weisen in mehreren

der vier Bereiche nahe beieinander liegende Werte auf, zeigen jedoch trotzdem auf mindestens einem Faktor nennenswerte Differenzen. Die empirische Bestimmung von fachspezifischen Unterschieden in den Einstellungsmustern der Studienanfänger*innen anhand eines faktorenanalytischen Vorgehens generiert damit einigen Interpretationsspielraum hinsichtlich der Unterscheidung von Fachkulturen. So kann argumentiert werden, dass nennenswerte Unterschiede zwischen einzelnen Fächern und Fächergruppen existieren, die deshalb unterschiedlichen Kulturen zugeordnet werden können. Dem kann entgegenhalten werden, dass einige dieser Fächer sich aber auch in mehreren Einstellungskomplexen stark ähneln und deshalb eine Kultur bilden, innerhalb derer allenfalls eine geringe Varianz an Einstellungen herrscht.

In Anlehnung an die Studie Windolfs versucht Armingeon Fachkulturen an der Universität Bern zu identifizieren. Er übernimmt dazu die Items zur Studienmotivation aus Windolfs Studie und führt ebenfalls Faktorenanalysen durch, um die Ergebnisse der Befragungen zu analysieren. Im Gegensatz zu Windolf umfasst die Untersuchung Armingeons allerdings nicht nur Studienanfänger*innen. Die mithilfe der Faktorenanalyse für die in dieser, teils andere Fächer umfassenden Stichprobe aufgezeigten Unterschiede in den abgefragten Einstellungen der Studierenden zwischen den berücksichtigten Fächern ähneln den Befunden Windolfs. Neben den Betriebswirtschaftler*innen fallen hier auch Mediziner*innen in die Kategorie karriereorientiert, sie weisen außerdem die höchsten negativen Werte für den Faktor Studium als Lebensstil auf. Die naturwissenschaftlichen Fächer liegen hier nicht wie bei Windolf im negativen Bereich. Auf dem Faktor Wissenschaftsorientierung laden die Volkswirtschaftler*innen nicht negativ, die Betriebswirtschaftler*innen weisen allerdings wie bei Windolf die höchsten Negativwerte auf, die Naturwissenschaftler*innen die höchsten Positivwerte. Benachbarte Fächer (Betriebswirtschaftslehre und Volkswirtschaftslehre, Politologie und Soziologie) unterscheiden sich auch hier in mehreren Faktoren in nennenswertem Maße (Armingeon 2001).

Georg untersucht den Einfluss verschiedener Studienmotive auf die Fachwahl in der zeitlichen Entwicklung genauer und vergleicht sie für drei Zeitpunkte (1985, 1995, 2004) mit dem Einfluss der sozialen Herkunft. Anhand des Konstanzer Studierendensurveys analysiert er hierzu, wie sich extrinsische und intrinsische Gründe für die Studienfachwahl, die Relevanz der drei Lebensbereiche Privates, Beruf und Studium sowie Politik und Kultur und extrinsische, intrinsische, altruistische, freizeitorientierte und wissenschaftsorientierte Berufsmotive auf die Studienfachwahl auswirken (Georg 2005).
Er stellt dabei anhand von zweitstufigen Logit-Modellen für die Motivmerkmale eine wesentlich höhere Erklärungskraft fest, als für die soziale Herkunft, die allerdings auch bei Kontrolle der Studienmotive für den Großteil der betrachteten Fächergruppen (Sozialwissenschaften, Kulturwissenschaften, Wirtschaftswissenschaften, Naturwissenschaften, Ingenieurswissenschaften, Jura) weiter eine

zwar schwache, aber signifikante Wirkung aufweist. Lediglich für Studierende der Ingenieurswissenschaften und der Rechtswissenschaften zeigt sich an einem bzw. an zwei der drei Messzeitpunkte kein Herkunftseffekt (im Vergleich zur Referenzkategorie der Medizinstudierenden). Für die einzelnen Fächergruppen zeigen sich auch bei Georgs Vorgehen deutliche Unterschiede hinsichtlich der Motive, die zur Aufnahme eines Studiums führen: Kulturwissenschaftler*innen streben eher eine wissenschaftliche Tätigkeit an, auch für Soziolog*innen ist diese Richtung ein Motiv. Dazu kommen ein ausgeprägtes intrinsisches Berufsinteresse sowie kulturelle Präferenzen. Auch bei den Wirtschaftswissenschaftler*innen finden sich Indizien für eine intrinsische Berufsmotivation, allerdings sind sie gleichzeitig deutlich stärker extrinsisch motiviert als Mediziner*innen. In den Naturwissenschaften fällt ein starkes Interesse an Freizeit und Wissenschaft auf. Dies bestätigt die Ergebnisse Windolfs und Armingeons, dass Studierende fachspezifische Motivationen für ihr Studium und auch für die darauf aufbauende Berufswahl aufweisen.

Die Befunde Windolfs, Armingeons und Georgs legen einen selektiven Prozess der Fachwahl nahe, in dessen Verlauf sich Studierende nicht nur entsprechend ihrer sozialen Herkunft, sondern auch gemäß ihrer Studienmotive und Überschneidungen mit bestimmten, innerhalb der Fachkulturen vorherrschenden Einstellungsmustern, auf die einzelnen Fächer verteilen. Mit Bezug auf Bourdieus Habituskonzept ist diese Erkenntnis nicht weiter verwunderlich, da die angehenden Studierenden ja bereits vor dem Eintritt in die Hochschule und damit zum Zeitpunkt der Fachwahl bereits einen (Klassen-) Habitus ausgebildet haben, der die Fachwahl vorstrukturiert.
Dass mit dem Beginn des Studiums dann eine fachspezifische Sozialisationsphase beginnt, die den Herkunftshabitus entsprechend der jeweiligen Fachwahl entweder ergänzt oder sogar überformt (Apel 1989; Borchert 1986; Frank 1990), widerlegen die Befunde der angeführten Studien nicht. Im Gegenteil dürfte eine in ihren soziodemographischen Merkmalen (Bülow-Schramm (2014:274) nennt es „studiengangspezifische Homogenität der Studierenden“) und in ihren Einstellungen und Verhaltensweisen bereits vor Eintritt in die Hochschule im jeweiligen Fach relativ homogene Studierendenschaft (für Germanistik und Betriebswirtschaftslehre: Gass/Meister 1996) günstige Umstände für den Prozess der fachlichen Sozialisation bieten. In jedem Fall jedoch weisen alle bisherigen Befunde darauf hin, dass die Ausbildung fachspezifischer Kulturmuster als Zusammenspiel von Selektion und Sozialisation verstanden werden muss (Huber 1991; Krais 1996).

Der Beitrag der deutschen Fachkulturforschung zur Analyse fachlicher Besonderheiten in epistemologischer, wissenschaftsorganisatorischer und kultureller Perspektive ist vor allem im Hinblick auf die theoretische Fundierung des Fachkulturkonzepts durch die Integration der Habitusforschung zu würdigen. Neue Erkenntnisse sind vor allem im Bereich der studentischen Fachkulturen, aber auch im

Hinblick auf ungleichheitsproduzierende Potentiale von Fachkulturen und in der Ausweitung der Analyse auf Selektionsprozesse produziert worden. Die potentiell für den Benotungsprozess relevanten untersuchten Differenzierungslinien sind jedoch zum größten Teil bereits in der US-Forschung benannt und untersucht worden. Die Ausnahme bilden drei Punkte: 1. Die Zusammensetzung der Studierenden in Bezug auf *leistungskonforme* Ursachen, mit möglichen Vorteilen für Akademiker*innenkinder und entsprechend zu erwartenden besseren Noten in Fächern (bzw. Studiengängen) mit einem hohen Anteil solcher Studierender. 2. Die ebenfalls leistungskonformen Auswirkungen von Unterschieden in der Lehrqualität, etwa aufgrund unterschiedlicher personeller Ausstattung 3. Fachspezifische Lernmotivationen, die theoretisch je nach Ausprägung sowohl bessere als auch schlechtere Noten begünstigen könnten.

Für die vergleichende Untersuchung der Notengebung verschiedener Fächer sind die ausgearbeiteten Unterschiede in den Benotungsbedingungen wichtige Ansatzpunkte der Analyse. Die Vielzahl der empirischen Studien zu unterschiedlichen Zeitpunkten seit den 1970er Jahren verdeutlicht, dass die untersuchten Distinktionskriterien langfristig wirksam sind. Dies lässt sich auch für die prüfungsrelevanten Merkmale zeigen:

Bereits 1973 erstellen Keil und Piontkowski anhand von Beobachtungen sowie Interviewaussagen von Studierenden und Dozierenden der Universität Münster charakteristische Beschreibungen der Lehr- und Lernumgebungen in verschiedenen Fächern. Sie unterscheiden dabei unter anderem stoffzentrierte von studierendenzentrierten Fächern und Fächer mit hohem Ordnungsgrad des Gelehrten von solchen mit niedrigem Ordnungsgrad. Mathematik, Biologie, Physik, Chemie, Geologie, Medizin, Wirtschaftswissenschaften, Anglistik, Romanistik, werden als stoffzentriert und mit hohem Ordnungsgrad beschrieben[26]. Auch die Lehre in den Rechtswissenschaften und in der Theologie wird als sachorientiert erfasst, hier werden allerdings keine Aussagen zum Ordnungsgrad getroffen. Geschichte wird als studierendenzentriert und stark geordnet beschrieben, Psychologie und Philosophie, Soziologie, Publizistik und Pädagogik als studierendenzentriert und mit geringem Ordnungsgrad (Keil/Piontkowski 1973). Für Medizin und Pädagogik scheinen diese Einordnungen auch heute nicht weniger treffend zu sein (Nierobisch 2010).

Portele lässt 11 Fächer, in denen Befragungen von Hochschullehrenden und Studierenden durchgeführt wurden, von Wissenschaftler*innen und Wissenschaftsadministrator*innen entsprechend dem

[26] Barz/Miethig (1993) befragen Professor*innen der Universität Kaiserslautern zu ihren Eindrücken der Lehre und Lehrorganisation. Auch dort werden die natur- und ingenieurswissenschaftlichen Fächer als stoffzentriert und in hohem Maße geordnet, zudem mit beschränkten Partizipationsmöglichkeiten für die Mitarbeiter*innen an den Lehrstühlen beschrieben

jeweiligen Grad der Standardisierung in eine Rangordnung bringen. Standardisierungsgrad wird dabei als „'Grad der Vorhersagbarkeit von Verhaltensweisen der Forschenden und Lehrenden auf Grund des Materials und der Kenntnis des Materials in dieser Disziplin'" (Portele 1975:107) verstanden. Mit steigendem Standardisierungsgrad stellt sich die Rangfolge der einbezogenen Fächer folgendermaßen dar: Politikwissenschaften, Soziologie, Germanistik, Anglistik, Archäologie, Betriebswirtschaftslehre, Rechtswissenschaften, Medizin, Mathematik, Chemie und Elektrotechnik[27]. Mit zunehmendem Standardisierungsgrad nehmen auch die „geregelten Verhaltensweisen" (ebd.:108), das Kontrollverhalten und die Zeit, die in Veranstaltungen und für das Lernen aufgewendet wird, in den Fächern zu. Die Vielfalt im Lehrangebot, interdisziplinäre Betätigung, und Partizipationsmöglichkeiten[28] für den Mittelbau und für Studierende sowie deren Wahlmöglichkeiten hinsichtlich des Lehrangebots nehmen hingegen mit steigender Standardisierung ab.

Die Organisation und Durchführung von Lehre wird in vielen Studien noch genauer untersucht, als im Hinblick auf ein übergreifendes Standardisierungsniveau - auch schon in den 1970er Jahren. Oehler et al. befragen Hochschullehrer*innen unterschiedlicher Hochschultypen der Region Frankfurt/Darmstadt in den Fächern Jura, Wirtschaftswissenschaften, Maschinenbau, Architektur Sozialpädagogik und in Lehramtsstudiengängen zu Zusammenhängen zwischen Problemen der Hochschuldidaktik und den Ausbildungskapazitäten von Hochschulen (Oehler et al. 1976 und 1978).
Die Ergebnisse ihrer Befragung offenbaren einige bereits angesprochene Unterschiede zwischen den erfassten Fächern, wie das Betreuungsverhältnis, das sich in der Architektur, der Sozialpädagogik, in den Rechtswissenschaften und im Maschinenbau wesentlich besser als in den Wirtschaftswissenschaften darstellt - wobei die Lehrenden in den Rechtswissenschaften die wenigsten Examenskandidat*innen im Vergleich zu den anderen Fächern kennen - gefolgt von den Wirtschaftswissenschaften an Universitäten. Lehrende im Maschinenbau kennen ihre Prüflinge hingegen zur Hälfte bzw. an Fachhochschulen sogar zu drei Vierteln, auch in Architektur und Sozialpädagogik kennen sie mindestens die Hälfte der Kandidat*innen persönlich. Auch das Zeitbudget für Forschung und Lehre sowie hier sogar schon explizit für die Prüfungstätigkeit wird bereits in den 1970ern als relevantes Unterscheidungskriterium erfasst: Sozialpädagogiklehrende verwenden mehr Zeit auf die Vorbereitung

[27] Bargel (1988) kommt bei einer Befragung von Studierenden aus 40 Einzelfächern zur wahrgenommen Höhe des Gliederungsgrad ihres Studiums zu einer ähnlichem Reihung. Dort zeigt sich außerdem ein positiver Zusammenhang zwischen Gliederungsgrad und Anforderungsniveau.
[28] Apenburg et al. (1977b) ermitteln die wahrgenommenen Partizipationsmöglichkeiten von Studierenden der Universität Saarbrücken mit ähnlichem Ergebnis. Ein Lehrstil mit hohen Partizipationsmöglichkeiten wird von den Befragten in den Fächern Französisch und Germanistik berichtet, über Medizin, Jura und Mathematik sinken die wahrgenommenen Partizipationsanreize hin zu Betriebswirtschaftslehre (bei Portele vor Jura, Mathematik und Medizin), Chemie und Physik.
Dippelhofer-Stiem (1983) kommt bei einer Analyse fachlicher Lehrumwelten zu ähnlichen Ergebnissen: Hohe Beteiligungsmöglichkeiten in Germanistik, niedriger in Medizin und am geringsten in den Natur- und Wirtschaftswissenschaften.

und Durchführung von Lehre als ihre Kolleg*innen aus anderen Fächern[29]. In Jura wird der höchste zeitliche Prüfungsaufwand in der Vorlesungszeit angegeben, gefolgt von den Wirtschaftswissenschaften und Maschinenbau an Fachhochschulen[30] (Oehler et al. 1978).

Teichler et al. führen in den 80er Jahren eine Längsschnittstudie durch, in deren Mittelpunkt Zusammenhänge zwischen den Studienerfahrungen von Absolvent*innen und deren Berufseinstieg stehen. Die Studie umfasst Befragungen von mehr als 1000 Prüflingen des Zeitraums 1983-1985 an 21 Hochschulstandorten. Zudem wurden an jeweils sieben Hochschulen für die drei Fächer Maschinenbau, Sozialarbeit/Sozialpädagogik und Wirtschaftswissenschaften verschiedene Elemente des Studienangebotes und der Studienbedingungen mittels Dokumentenanalysen sowie anhand von Leitfadeninterviews mit Angehörigen unterschiedlicher Hochschulbereiche erfasst.

Als charakteristisch für Maschinenbau werden eine hoch standardisierte Wissensvermittlung, ein hoher Zeitdruck und die Vermittlung instrumenteller Problemlösungskompetenzen betrachtet. Es herrscht im Maschinenbau die Überzeugung, objektiv geteilte Qualitätsstandards bilden die Basis für gegenseitige Beurteilungen, die Studierenden versuchen in Lerngruppen gemeinsam das hohe Lernpensum zu bewältigen. In der Sozialarbeit hingegen werden die Anforderungen für überschaubar gehalten. Es wird Wert darauf gelegt, möglichst vielen Studierenden ein erfolgreiches Studium zu ermöglichen. Hier zeigt sich, dass fachspezifische Selektionsneigungen nicht nur als Folge fachspezifischer Zusammensetzungen von Lehrenden, etwa in Hinblick auf die Wahrnehmung von Berufschancen oder die Entwicklung des Studierendenengagements im Zeitverlauf[31] und als Nebeneffekt von generell eher liberaleren pädagogischen Werten entstehen können. Offensichtlich wird zumindest in einzelnen Fächern offen reflektiert, wie stark die Selektion im Prüfungsprozess ist bzw. wie stark sie sein sollte.

In den Wirtschaftswissenschaften zeigen sich ebenfalls stark standardisierte Inhalte und Lehrmethoden. Hier wird außerdem der geringe Kontakt zwischen Lehrenden und Lernenden außerhalb von

[29] Noelle-Neumann ermittelt für Geisteswissenschaftler*innen den größten Zeitaufwand für die Vorbereitung von Lehre, für Mediziner*innen den geringsten (Noelle-Neumann 1980).

[30] Hinsichtlich der Relation Prüflinge pro Prüfer*in finden sich unterschiedliche Befunde (allerdings auch zu unterschiedlichen Zeitpunkten): Bei Noelle-Neumann findet sich hinsichtlich der Prüfungsbelastung auf Fächergruppenebene eine Abstufung von den Rechts-, Sozial- und Wirtschaftswissenschaften mit der höchsten Zahl an Prüflingen pro Hochschullehrer*in über die Naturwissenschaften hin zu den Geisteswissenschaften mit den durchschnittlich wenigsten Prüfungskandidat*innen (Noelle-Neumann 1980:147). Von Kopp und Weiß (1995) hingegen finden in den Sprach- und Kulturwissenschaften die meisten Prüflinge pro Kopf, in Mathematik und den Naturwissenschaften die wenigsten.

[31] Siehe hierzu Enders/Teichler (1995): Vor allem wirtschaftswissenschaftliche und ingenieurswissenschaftliche Lehrende, gefolgt von denen der Erziehungswissenschaften sind der Ansicht, dass die Studierenden fünf Jahre vor der Befragung besser auf Veranstaltungen vorbereitet waren als die zum Zeitpunkt der Befragung Studierenden. In den Kunst- und Rechtswissenschaften sowie in der Medizin wird diese Ansicht am seltensten vertreten. Umgekehrt sind Lehrende in den Ingenieurs- und Rechtswissenschaften am seltensten der Meinung, dass die Studierenden besser vorbereitet seien, als noch fünf Jahre zuvor, in den Kunstwissenschaften findet sich diese Ansicht am häufigsten.

Lehrveranstaltungen[32] und die Überfüllung der Veranstaltungen[33] hervorgehoben (Teichler et al. 1987). Letzteres könnte einen Einfluss auf die Notengebung haben, wenn davon ausgegangen wird, dass die Leistung der Studierenden bei überfüllten Veranstaltungen aufgrund der schlechteren Betreuungsrelation sinkt. Umgekehrt könnte aber auch ein Wiedergutmachungseffekt eintreten, wenn Prüfende veranlasst werden, gute Noten als Ausgleich für die schlechten Bedingungen zu geben (Müller-Benedict/Tsarouha 2011).

Quantifizierbare Unterschiede erarbeiten Teichler et al. für den Grad der Strukturierung des Studiums. Hier zeigt sich, dass von den Studierenden in Maschinenbau deutlich mehr Zeitaufwendung für den Besuch von Lehrveranstaltungen erwartet wird als von den Studierenden der anderen beiden Fächer und ihr Studium auch einen deutlich höheren Anteil an Pflichtveranstaltungen aufweist. In Kontrast dazu ist das Beratungsangebot für Studierende des Maschinenbaus (wie auch der Wirtschaftswissenschaften) eher gering, das in der Sozialarbeit hingegen sehr hoch. Die Prüfungsbelastung in den Wirtschaftswissenschaften und im Maschinenbau wird als studienbegleitend beschrieben, während sie in der Sozialarbeit eher als punktuell eingeschätzt wird. Während im Maschinenbau zudem auch an der eigenen Hochschule eher anonym studiert wird, werden in der Sozialarbeit hochschulübergreifende Kontakte gepflegt. Private Kontakte zwischen Studierenden und Hochschullehrenden sind eher selten, kommen am ehesten in der Sozialarbeit vor. Nur unter den Sozialarbeitsstudierenden werden auch ein gemeinsamer Lebensstil und politische Aktivitäten als Basis des Kontakts zu Lehrenden genannt.

Befragt zu den Leistungsanforderungen und der Informiertheit über Prüfungserwartungen seitens der Hochschullehrenden weisen die Studierenden kaum fachspezifische Unterschiede in der wahrgenommenen Transparenz der Erwartungen auf[34], sind sich jedoch dem spezifischen Typ der Wissensvermittlung in ihrem Fach bewusst: Studierende der Wirtschaftswissenschaften und in Maschinenbau sehen die Wiedergabe des Gelernten als wesentliches Ziel der Prüfungen, die der Sozialpädagogik hingegen betonen, dass von ihnen vor allem Problemlösungskompetenz gefordert wird. Während die

[32] Generell liegen die Wirtschaftswissenschaften hinsichtlich der Kontakthäufigkeit zwischen Lehrenden und Studierenden in allen Studien stets im hinteren Bereich (Gawatz, 1991; Multrus et al., 2008; Ramm et al, 2011: Der meiste Kontakt findet in den Kulturwissenschaften statt, es folgen Sozial-, Natur-, Ingenieurswissenschaften und Medizin - je nach Untersuchung in unterschiedlicher Reihenfolge - und schließlich finden sich am Ende stets die Wirtschaftswissenschaften und Jura). Nicht weiter verwunderlich ist da, dass Studierende der Betriebswirtschaften von einem schlechten Verhältnis zu ihren Dozierenden berichten (Apenburg et al. 1977b)

[33] Auch Enders und Teichler (1995) berichten von hohen Teilnehmer*innenzahlen in wirtschaftswissenschaftlichen Veranstaltungen. Eine ebenfalls starke Nachfrage findet sich in Jura und in den Ingenieurswissenschaften (ebd.; Oehler et al. 1978; Portele 1976 - Letzterer berichtet zudem von hohen Teilnehmer*innenzahlen in Medizin).

[34] Bei Apenburg et al. (1977b) findet sich eine Abstufung der Informiertheit über Prüfungsanforderungen von den sich am besten informiert gefühlten Studierenden in Jura über Mathematik, Germanistik, Physik und Romanistik hin zu Medizin und Chemie, sowie zur Betriebswirtschaftslehre, wo sich die Studierenden am wenigsten klar über die an sie gestellten Anforderungen in den Prüfungen waren. Dass die Betriebswirtschaftslehre am Ende der Rangliste steht, verwundert nicht weiter, scheint dort auch ein besonders geringes Interesse daran zu bestehen, Prüfungsangelegenheiten mit den Lehrenden zu besprechen (Gleich et al. 1982:120)

ersten beiden Gruppen auch ihre Belastbarkeit unter Prüfungsbedingungen als im Rahmen von Prüfungen evaluiert sehen, spielt dieser Aspekt in der Sozialarbeit keine große Rolle.

Nicht nur der Standardisierungsgrad des gelehrten und geprüften Wissens, auch das Verhältnis von Lehre und Forschung sowie die Betreuungsrelation und die Art des Kontakts zwischen Studierenden und Lehrenden wurden in Deutschland demnach schon vor Etablierung der expliziten Fachkulturforschung untersucht und dienen damit seit mehreren Jahrzehnten als zentrales Charakterisierungsmerkmal von Fächern. Dass die Beschäftigung mit solchen Charakteristika in der Forschung jahrzehntelang andauert, ist ein Beleg für die langjährige Existenz dieser Differenzierungslinien. Führen entsprechende Unterschiede zu unterschiedlichen Benotungsmustern, bedeutet dies, dass sich Differenzen in der Benotung über einen langen Zeitraum nachweisen lassen sollten. Bei Teichler et al. findet sich mit der Form der Prüfungsbelastung (punktuell vs. studienbegleitend) zudem ein empirischer Beleg für die von Huber angeführten Unterschiede in der Prüfungsgestaltung der Fächer. Interessant ist, dass sich der Grad der Informiertheit über Prüfungserwartungen bei den Studierenden entweder im Zeitverlauf zwischen den Fächern angleicht oder sich in der 10 Jahre zuvor durchgeführten Befragung von Apenburg et al. (1977) eine hochschulspezifische Abweichung abzeichnet.

Dass sich die tatsächlichen Prüfungsbedingungen in der Tat von Fach zu Fach verschieden darstellen, weisen neben Teichler et al. auch Helberger und Schulz nach. Deren Analyse des Einflusses formaler Prüfungsbedingungen auf die Studiendauer in den Fächern Wirtschaftswissenschaften, Mathematik und Jura zeigen, dass die Bandbreite formaler Regelungen in den Diplomprüfungen der Wirtschaftswissenschaften und in Mathematik deutlich größer ist als in den Staatsexamina (SE) der Rechtswissenschaften. Die größere Varianz der Prüfungsbedingungen der Diplomstudiengänge erklärt die dort festgestellte längere Studiendauer gegenüber dem Staatsexamen in Jura (Helberger/Schulz 1987). Wird eine längere Studiendauer parallel zu schlechteren Noten als geringerer Bildungserfolg verstanden (etwa Krempkow 2006; Mosler/Savine 2004; Wittenberg 2005), ließen sich als Folge heterogenerer Prüfungsbedingungen auch schlechtere Abschlüsse vorstellen. Als Mechanismus käme dann eine größere Unsicherheit über die Prüfungserwartungen in Frage, die eine weniger gezielte Vorbereitung zur Folge hätte. Unabhängig vom Notenniveau ist bei einer höheren Bandbreite an formalen Regelungen zudem mit einer größeren Streuung der Noten zu rechnen. Inwiefern punktuelle oder studienbegleitende Prüfungsbelastung unterschiedliche Resultate begünstigen, dürfte von der Art der Wissensabfrage abhängen. Hohe punktuelle Belastungen dürften vor allem bei der Abfrage von Faktenwissen schwieriger zu bewältigen sein, als kleinere, häufiger anstehende Überprüfungen. Bei vornehmlich interpretatorischen Leistungen sollte kein großer Unterschied bestehen.

Enders und Teichler (1995) zeigen, dass sich auch die Art der für den erfolgreichen Abschluss einer Lehrveranstaltung erforderlichen Leistungen fachspezifisch darstellt: Während in den Natur- und

Ingenieurswissenschaften im Vergleich zu den anderen Fächergruppen häufiger mehrere kleine schriftliche Leistungen und seltener eine große schriftliche Arbeit erwartetet werden, ist dies in den Sprach-, Kultur- und Kunstwissenschaften und am deutlichsten in den Sozialwissenschaften anders herum. In diesen Fächergruppen und in den Erziehungswissenschaften wird auch der größte Wert auf ein mündliches Referat gelegt, welches in den Ingenieurs- und Rechtswissenschaften wie auch in der Medizin weniger Gewicht besitzt. In den Erziehungs- und Rechtswissenschaften ist eine ähnliche, wenn auch nicht ganz so starke Bevorzugung großer schriftlicher gegenüber kleiner schriftlicher Arbeiten zu erkennen. In den Wirtschaftswissenschaften gibt es keine Präferenz für eine der beiden Arten, wie (noch stärker) in der Medizin ist die Rolle der schriftlichen Arbeiten relativ unbedeutend. In den wirtschafts- und Ingenieurswissenschaften wird eine Abschlussprüfung als zentrale Leistung für den Abschluss einer Lehrveranstaltung erwartet, in den Sozial- und den Sprach- und Kulturwissenschaften ist eine solche Leistung nur selten von Bedeutung. Wie zu Beginn der Arbeit erläutert, können diese Unterschiede die Notengebung dadurch beeinflussen, dass Leistung unterschiedlich genau abgefragt wird, bei schriftlichen Hausarbeiten zudem unabhängig davon bessere Noten zu erwarten sind.

Zusammenfassend lässt sich festhalten, dass sich eine Vielzahl von Merkmalen finden lässt, anhand derer sich Fächer unterscheiden lassen. Die Anwendung zahlreicher Distinktionslinien lässt sich wie gezeigt über einen langen Zeitraum zurückverfolgen, was die langfristige Existenz fachlicher bzw. fachkultureller Grenzziehungen belegt. Entsprechend ist, wird von einem Einfluss der Fachkulturen auf die Notengebung ausgegangen, mit zeitlich stabilen fachspezifischen Mustern der Notengebung zu rechnen. Da die konkreten Ausprägungen fachspezifischer Prüfungsbedingungen nicht als Konstanten wirken, muss dabei berücksichtigt werden, dass sie sich von Fach zu Fach in unterschiedlichem Maße und zu unterschiedlichen Zeitpunkten verändern. Solchen fachspezifischen Entwicklungen gemäß ist anzunehmen, dass die Unterschiede in den Prüfungsbedingungen und damit auch in den Notenmustern, wenn auch über eine bestimmte Zeitspanne stabil, nicht immerwährend gleich bleiben, sondern vielmehr fachspezifische Entwicklungen zu erwarten sind.

F**H2a: An deutschen Hochschulen existieren hochschulübergreifend zeitlich stabile Differenzen im Notenniveau zwischen fachlich abgegrenzten Studiengängen**

FH2b: Zeitlich stabile Differenzen im Notenniveau zwischen fachlich abgegrenzten Studiengängen lassen sich zum Teil auf fachspezifische Prüfungsbedingungen zurückführen

F**H3a: An deutschen Hochschulen existieren hochschulübergreifend fach-/studiengangspezifische Notenverläufe**

FH3b: Fach-/studiengangspezifische Notenverläufe lassen sich zum Teil auf fachspezifische Entwicklungen der Prüfungsbedingungen zurückführen

Für die Identifizierung prüfungsrelevanter Unterschiede ist es prinzipiell unerheblich, ob sich die Unterscheidungsmerkmale im Grunde alle auf epistemologische und wissenschaftsorganisatorische Charakteristika und Strukturen zurückführen lassen oder zum Teil genuin kulturelle Komponenten darstellen, solange die vermuteten Effekte theoretisch begründbar sind. Wird die Dreiteilung der Fächerdistinktion als analytisches Schema akzeptiert, ergibt sich aus den drei Kategorien folgendes Muster vermuteter Wirkungszusammenhänge:

- *Epistemologische Merkmale,* wie die Existenz eines einheitlichen Paradigmas / eines einheitlichen Wissensfundus oder die Struktur der Wissensakkumulation, bergen über den *Grad der Standardisierung* des vermittelten Wissens und dessen Abfrage in der Prüfung vor allem einen möglichen Einfluss auf die *Leistungsmessung*.

- *Wissenschaftsorganisatorische Merkmale*, wie die Ausrichtung des Fachs als rein vs. angewandt, seine Beziehungen zur Restgesellschaft oder das Verhältnis von Forschung und Lehre, beeinflussen am ehesten das *Selektionsklima* und damit die *Beurteilung* der Leistung[35].

- *Kulturelle Merkmale*, wie die Zusammensetzung der Studierenden und der Lehrenden beeinflussen die Notengebung möglicherweise sowohl durch *leistungskonforme Auswirkungen* auf die *Leistungsmessung* als auch über das *Selektionsklima* auf die *Beurteilung* der Leistung.

Dieses Zusammenhangsmuster ergibt sich aus den zuvor beschriebenen potentiellen Mechanismen der Beeinflussung des Benotungsprozesses. Es ist als idealtypische Konstruktion zu verstehen und die einzelnen Zusammenhänge müssen als theoretisch hergeleitete Annahmen angemessenen empirischen Überprüfungen standhalten.

[35] Natürlich kann das Selektionsklima auch schon vor der eigentlichen Beurteilung auf die Notengebung einwirken, wenn in Prüfungen (bewusst) leichte bzw. schwere Aufgaben gestellt werden. Den Akt der Messung selbst beeinflusst dies streng genommen jedoch nicht. Vielmehr handelt es sich in solchen Fällen um einen vorgezogenen Einfluss auf die Beurteilung.

4. Hochschulspezifische Bedingungen der Notengebung

Auch innerhalb von Fächern und Studiengängen existieren an verschiedenen Hochschulen unterschiedliche Prüfungsbedingungen, die hochschulspezifische Notenniveaus begünstigen. Dies wird schon beim formalen Rahmen der Prüfungen deutlich, da die Prüfungsordnungen, mit Ausnahme der landeseinheitlichen Staatsexamina, von den einzelnen Hochschulen ausgearbeitet werden und nur dort Gültigkeit besitzen. Aufgrund der hierzulande stärkeren Identifikation mit der Fachkultur als mit der beschäftigenden Hochschule (Enders/Teichler 1995) kann davon ausgegangen werden, dass nicht-formale Prüfungsbedingungen, die auf die Wahrnehmungs-, Denk- und Handlungsmuster der Prüfenden wirken, vor allem im Rahmen der fachspezifischen Sozialisation Einfluss erhalten. Abschlussspezifische Unterschiede sollten sich vor allem durch formale Unterschiede im Prüfungsprozess ergeben, da keine speziellen „Lehramtskulturen", „Magisterkulturen" oder „Diplomkulturen" bekannt sind und nach heutigem Stand des Wissens davon ausgegangen werden muss, dass sich die nicht-formalen Prüfungsbedingungen abschlussübergreifend nach fachlicher Zugehörigkeit verteilen. Diese Unterteilung sollte jedoch als tendenziell verstanden werden.

Dass nicht-formalen Prüfungsbedingungen ein vergleichsweise starker Einfluss auf Fachebene zugeschrieben werden kann, bedeutet nicht, dass sie eine Art fachspezifischen, deterministisch wirkenden ‚Fraktionszwang' ausüben. Es muss in Betracht gezogen werden, dass immer wieder individuelle, von den im fachlichen Prüfungskontext üblichen Normen und Werten abweichende Einstellungen und Handlungen von Prüfenden existieren, die innerhalb der vorherrschenden Prüfungssysteme wirken. Sie können bei entsprechender Dynamik auch zu alternativen hochschulspezifischen wie hochschulübergreifenden Prüfungssystemen wachsen. Versteht sich etwa der Fachbereich einer Hochschule als besonders im Vergleich zu den gleichen Fachbereichen an anderen Hochschulen, zum Beispiel aufgrund einer langen, erfolgreichen Tradition oder aufgrund von besonderen Erfolgen bei der Drittmitteleinwerbung (Stichwort ‚Exzellenz'), könnte dies sowohl eine vergleichsweise schwache Selektion (‚wer hier studiert, ist exzellenter Lehre ausgesetzt und erbringt damit zwangsläufig gute Leistungen') als auch eine vergleichsweise starke Selektion (‚wer hier studiert, muss besonders hohen Ansprüchen genügen und beweisen, dass ihm dies gelingt') in Prüfungen begünstigen. Als einer schwachen Selektion zugrundeliegender Mechanismus wäre denkbar, dass Lehrende den Erfolg - und auch den Misserfolg - ihrer Studierenden ihrer eigenen Lehrleistung zuschreiben und sich in guten Noten selbst bestätigt sehen (vgl. Mansfield 2001). Für die Gegenthese der härteren Selektion spricht, dass Lehrende, die ihre Bestätigung bereits aus einer hohen fachlichen Reputation ziehen können, dafür nicht auf Prüfungsergebnisse zurückgreifen müssen. Ein solches Verhalten passt eher zu Lehrenden, die über keine andere Quelle der Bestätigung verfügen und deshalb ein hohes Bedürfnis an sozialer Anerkennung unter den Studierenden aufweisen (Crowl 1984).

Sollten traditionelle Reputationsmechanismen für eine Abweichung vom Fachdurchschnitt verantwortlich sein, ist davon auszugehen, dass diese Abweichung im Zeitverlauf recht stabil bleibt, da die Darstellung von Reputation unabhängig von hochschulinternen Veränderungsprozessen immer im Interesse des Fachbereichs bzw. der ganzen Hochschule ist (Müller-Benedict/Tsarouha 2011). Andere Mechanismen der hochschulspezifischen Abgrenzung von im Fach üblichen Benotungsrichtlinien können dagegen auch zu Brüchen in der Notenentwicklung von Hochschulen führen, wenn Fachbereiche, Fakultäten oder Institute neu strukturiert werden, das Lehrpersonal generationenintern oder generationenwechselnd ausgetauscht wird (vgl. ebd.). Auch wenn aufgrund fachbereichsinterner Kommunikation davon ausgegangen werden kann, dass sich die Bewertungsstandards und damit auch das Notenniveau der einzelnen Lehrenden aneinander angleichen (ebd.), dürften Personalwechsel in kleinen Fachbereichen/Instituten mit einer geringen Anzahl an Lehrenden eher Bruchstellen in Bezug auf einen konstanten Notenverlauf bieten als in großen, in denen ein einzelner Wechsel die Verhältnisse im Bestand des Lehrpersonals nur geringfügig ändert.

4.1 Hochschulen als „besondere Organisationen" einer funktional differenzierten Gesellschaft

Dass es nicht ungewöhnlich ist, dass sich die Notengebung zwischen Hochschulen innerhalb des gleichen Fachs unterscheidet, kann aus gesellschaftstheoretischer Perspektive, unter Einbeziehung organisationsanalytischer Erkenntnisse, genauer erklärt werden. Hochschulen als Organisationen erfüllen in einer funktional differenzierten Gesellschaft zwei wesentliche Funktionen: Sie produzieren (und vermitteln) Wissen und sie qualifizieren und selektieren Arbeitskräfte. Die Selektion erfolgt wie bereits beschrieben anhand der Überprüfung des Gelehrten und der anschließenden Benotung. Aus organisationstheoretischer Perspektive entsteht dabei ein Bild, das nur wenig Anzeichen einer irgendwie gearteten Stabilität in der Notengebung an Hochschulen aufweist: Im Fokus des wissenschaftlichen Betriebes der Organisation Hochschule steht das wissenschaftliche Personal. Dieses führt Forschung und Lehre als Mittel zur Erreichung der übergeordneten „output goals" (Gross 1968:522) Wissensproduktion und Arbeitskräfteselektion durch, ist aber auch in administrative Aufgaben eingebunden. Daraus ergibt sich eine spezielle Berufssituation, in der Hochschullehrende über die Möglichkeit verfügen „to ensure that he retains some control over the decisions that affect his work" (Mintzberg 1979:358). Vor Einsetzen der New Public Management-Welle werden Hochschulen als „Professional Bureaucracy" (ebd.), ausgestattet mit einem mächtigen Kern von spezialisierten Mitarbeiter*innen oder als „organisierte Anarchien" (Cohen et al. 1972:1) mit unklaren Verfahrensweisen, ohne Verständnis für die intern ablaufenden, im Versuchs- und Irrtumsverfahren generierten Prozesse beschrieben. Sie gelten als „organisationelle Hülle" (Huber 2005:394) ohne exklusive Gestalt

oder als “unvollständige Organisationen“ (Brunsson/ Sahlin-Andersson 2000:722), die unter anderem durch nicht ausreichende Kontrollmechanismen gekennzeichnet sind.

Inmitten dieser Rahmenbedingungen scheinbarer Willkür ohne greifende Kontrolle fällen Hochschullehrende in ihrer Rolle als Gatekeeper mittels Prüfung und Bewertung zentrale Urteile über die „Übergänge im Lebensverlauf“ (Struck 2001:31) von Studierenden. Werden diese allgemeinen Beschreibungen der Entscheidungsfindung bzw. der Prozessanwendung in Hochschulen auf die Notenvergabe übertragen, ergibt sich ein Bild, in dem keine einheitlichen Bewertungsverfahren existieren, die angewandten Verfahren eher experimentell entstanden sind, die zugrundeliegenden Mechanismen im Zweifelsfall nicht einmal verstanden werden und mangelnde Kontrolle über den gesamten Notengebungsprozess diesen Zustand perpetuiert.

Wenn auch etwas übersteigert, hebt dieses Bild den wesentlichen Argumentationspunkt hervor, der aus der organisationstheoretischen Perspektive auf die Universität in Bezug auf die Notengebung folgt: Die unvollständige Organisation Hochschule überträgt mangels struktureller Rahmung die Verantwortung für den Prozess der Notengebung, abgesehen von einigen formalen Vorgaben zum Prüfungsablauf und zum Format der Bewertung, weitestgehend auf die einzelnen Hochschullehrenden. Dass deren Bewertung dennoch nicht vollkommen subjektiv verläuft, sondern in einen fachlichen Kontext eingebettet ist, wurde im letzten Kapitel ausführlich dargestellt. Es ist dabei aus kulturtheoretischer Sicht mit fachspezifisch stabilen Benotungsmustern zu rechnen, da die Prüfenden die Messung und Bewertung im Rahmen ihrer fachspezifischen Sozialisation erlernen.

Wie im Rahmen dieser Ausführungen schon angeklungen ist, erfolgt die fachspezifische Sozialisation jedoch nicht nur im Rahmen der nationalen oder internationalen Fachgemeinschaft sondern vor allem im lokalen Kontext der hochschulspezifischen Arbeitsumgebung. Da Hochschulen keiner einheitlichen Ausgestaltung als Organisation unterliegen, ergibt sich die Möglichkeit hochschulspezifischer Abweichungen von fachspezifischen Prüfungs- und Bewertungsmustern, die ihrerseits räumlich wie zeitlich lokale Stabilität aufweisen sollten, falls sie sich ausbilden. Prüfende passen sich diesen Überlegungen nach den im jeweiligen Fachbereich/Institut herrschenden Praktiken der Notengebung an und reproduzieren die dort herrschenden Bewertungsstandards, wodurch auch Unterschiede in der Notengebung zwischen Hochschulen im gleichen Fach ermöglicht werden. Dass die hochschulspezifischen Benotungsmuster auch mittel- oder langfristig von den fachspezifisch etablierten Mustern abweichen, könnte entsprechend der dargelegten Argumentation auf die organisationale Struktur von Hochschulen zurückgeführt werden: Einmal durch lokale Einflüsse an den einzelnen Hochschulen in unterschiedlichen Versuch- und Irrtumsverfahren entstandene Bewertungspraktiken entwickeln eine Eigendynamik, indem sie aufgrund mangelnder Kontrollmechanismen von Prüfer*in zu Prüfer*in weitergegeben werden, wodurch sich das jeweils gegebene Notenniveau reproduziert.

Fraglich ist jedoch, inwieweit sich die allgemein für universitäre Prozesse diagnostizierten Vorgaben- und Kontrollmängel tatsächlich auf den Notengebungsprozess übertragen lassen. Fallen die Überprüfung von Gelehrtem und die Bewertung des Überprüften mit in die Kategorie weitestgehend willkürlich ablaufender Prozesse der Entscheidungsfindung und Prozessanwendung an Hochschulen? Oder wird die Willkür der einzelnen Prüfenden nicht nur durch fachspezifische und hochschulinterne Sozialisationsprozesse, sondern auch durch hochschulspezifische Rahmenbedingungen des Prüfungs- und Bewertungsprozesses eingeschränkt? Sollte letzteres der Fall sein, müssten sich mögliche Unterschiede in der Notengebung zwischen Hochschulen durch Unterschiede in den lokalen Prüfungs- und Bewertungsbedingungen erklären lassen.

4.2 Systemische Kopplung und externe Einflüsse

Wann immer über hochschulinterne und hochschulspezifische Entscheidungsprozesse gesprochen wird, muss berücksichtigt werden, dass Hochschulen als Organisationen im wissenschaftlichen Funktionssystem nicht unabhängig von dessen Umwelt operieren (vgl. Luhmann 1975; 1992). Sowohl durch die Funktion der Produktion von hochqualifizierten Arbeitskräften als auch durch die Notwendigkeit, Finanzmittel zur Wissensproduktion zu erhalten, ist das Wissenschaftssystem eng an das Wirtschaftssystem und an das politische System (welches wiederum inzwischen vom Wirtschaftssystem abhängig ist) gekoppelt. Es geht dabei seit Jahrzehnten ein sich stetig erhöhender Ökonomisierungsdruck vom Wirtschaftssystem aus (Schimank 2013). Nicht nur die generelle finanzielle Ausstattung und mit dieser zusammenhängende Merkmale wie Lehrqualität werden auf diese Weise extern beeinflusst.

Durch die zunehmende Evaluierung aller irgendwie messbaren Indikatoren für die von Hochschulen erbrachten Forschungs- und Ausbildungsleistungen (Stichwort „Evaluitis“ (Frey, 2008:125)) gerät auch die Notengebung selbst immer wieder in den Fokus, wenn es darum geht, die Lehrleistung von Dozierenden, Instituten, Fakultäten oder auch einer ganzen Hochschule zu bewerten. Der Selektionsprozess an Hochschulen, der in Form der Prüfung stattfindet, produziert also Werte, die sich für verschiedene Analyseebenen messen lassen. Noten auf der Aggregatebene werden dadurch ebenso als Indikator nutzbar, wie die Noten des einzelnen Prüflings. Nicht nur dessen Noten werden als Abbild von Leistung interpretiert, auch Noten im Aggregat werden auf Hochschulebene, vor allem im angelsächsischen Bildungsraum, aber auch zunehmend in Europa, als Indikator für die Ausbildungsleistung der einzelnen Einrichtungen genutzt und zum Zwecke der Wettbewerbsförderung mit dem Erhalt von Fördermitteln verknüpft (zu diesem Zusammenhang siehe etwa Bagues et al. 2008). Indirekt stehen die Noten auch mit der Zahl der Absolvent*innen in Zusammenhang, einem Outputindikator, der auch in Deutschland bereits zur Mittelvergabe genutzt wird, wenn auch in bisher vergleichsweise geringem Umfang (Bauer/Grave 2011). Der Output des Notengebungsprozesses, die darin vergebe-

nen Noten, bietet entsprechend (wie Leistungsindikatoren grundsätzlich – vgl. Frey, 2008) einen Ansatzpunkt, um die Informationen die mit ihm verknüpft werden, zu beeinflussen. Kritische Stimmen weisen in der Diskussion um die Nutzung von Ergebnisindikatoren immer wieder auf den mit der Wahl von Noten als Indikator verbundenen Manipulationsanreiz hin (etwa Weyer 2013).

Die Abhängigkeit vom Wirtschaftssystem kann jedoch neben der Möglichkeit einer intentionalen Beeinflussung des Outcomes auf Aggregatebene auch weitere, nicht-intendierte Auswirkungen auf das durchschnittliche Notenniveau bestimmter Untersuchungseinheiten produzieren. Da die Produktion von hochqualifizierten, in der Regel fachspezifisch ausgebildeten Arbeitskräften - einerseits vom Angebot der Hochschulen (beeinflusst etwa durch demographischen Wandel), andererseits vom Bedarf des Wirtschaftssystems (beeinflusst etwa durch technologischen Wandel) abhängig - immer an die Aufnahmefähigkeit des Arbeitsmarktes, sowohl generell, als auch in fachspezifischen Sektoren gekoppelt ist (Müller-Benedict 2005; Titze 1990), ist auch die Selektion innerhalb des Angebots, also innerhalb der Gesamtheit der Absolvent*innen, abhängig vom Bedarf: Je größer er ist, umso geringer wiegt das in der Abschlussnote festgehaltene Ergebnis der Ausbildung im Vergleich zur Tatsache, dass sie überhaupt erfolgreich absolviert wurde. Je geringer der Bedarf allerdings, umso eher dürfte die Abschlussnote als Indikator für die Leistung, mit der der Prüfling seine Ausbildung beendet hat über seine Chancen auf dem Arbeitsmarkt entscheiden.

Nath/Dartenne/Oelerich (2004) konnten in Längsschnittanalysen der Konjunkturen von Lehrer*innen zeigen, dass es kein über die Zeit stabiles Selektionsklima von Lehrer*innen gibt, sondern die Selektionsschärfe auch durch die zahlenmäßige Entwicklung des Bildungssystems (mit-)bestimmt wird. Sollten die Prüfenden als diejenigen, die zugespitzt formuliert mit ihrem Urteil die Entscheidung über das Schicksal ihrer Studierenden fällen, in irgendeiner Weise auf diese Abhängigkeit vom Arbeitsmarkt reagieren, sei es durch besondere Milde oder im Gegenteil durch besondere Strenge (Hitpass/Trosien 1987), könnten sich Überfüllung und Mangel auf dem Arbeitsmarkt auf diese Weise auf das Notenniveau ausüben (Müller-Benedict/Tsarouha 2011). Bei der Analyse entsprechender Entwicklungen sollte dabei wohlgemerkt die fachspezifische Arbeitsmarktlage berücksichtigt werden (Reisz/Stock 2013). Hochschulspezifische Abweichungen sind nur in Studiengängen zu erwarten, die nicht auf einen spezifischen Beruf vorbereiten und keinen einheitlichen Arbeitsmarkt bedienen, wie etwa die Germanistik.

Durch die Existenz funktionaler Verzerrungsanreize wie systemischer Abhängigkeiten können Noten also sowohl den Charakter eines strategischen Instruments als auch den einer unintendierten Folge sozialen Handelns erhalten, was ihre Zuverlässigkeit in Hinblick auf ihre primäre Funktion auf der Individualebene, der Abbildung von Leistung, gefährdet. Bewusst oder unbewusst vollzogene „Manipulationen“ des Durchschnittsniveaus können jedoch, sollten sie existieren, nicht oder höchstens in

Einzelfällen nachträglich vollzogen werden, sie müssen in den individuellen Prüfungs- oder Bewertungsprozess integriert werden. Sie stellen dementsprechend Einflüsse auf das Selektionsklima der Prüfenden dar. Die eigentliche Verzerrung der Leistungsabbildung vollzieht sich dann in der Regel entweder im Prüfungsprozess, etwa durch das Stellen besonders leichter Klausuren, oder im Bewertungsprozess, durch besonders milde oder besonders strenge Beurteilung.

Aus den in diesem Kapitel ausgeführten theoretischen Überlegungen lassen sich folgende Forschungshypothesen ableiten:

FH4a: An einzelnen Hochschulen existieren signifikante, im Zeitverlauf stabile Abweichungen vom durchschnittlichen Notenniveau im jeweiligen Studiengang

FH4b: Signifikante, im Zeitverlauf stabile Abweichungen vom durchschnittlichen Notenniveau im jeweiligen Studiengang an einzelnen Hochschulen sind zum Teil auf hochschulspezifische Ausprägungen der Prüfungsbedingungen zurückzuführen

FH5: Es existieren hochschulinterne Einflüsse auf die Notengebung, die Brüche im Notenniveau einer konkreten Hochschule produzieren und es im Zeitverlauf verändern

FH6: Je nach Existenz eines einheitlichen Arbeitsmarkts verlaufen Notenniveaus studiengang- oder hochschulspezifisch in Abhängigkeit von Entwicklungen des Wirtschaftssystems

5. Zwischenfazit

Dass die Notengebung als Beurteilungspraxis anfällig für verschiedene Verzerrungen ist, ist aus der Schulforschung bestens bekannt. Noten bilden nicht nur die abgefragte Leistung ab, sondern beinhalten zusätzlich die Effekte von Messfehlern, individuellen Beurteilungsmaßstäben und Fehlern in der Beurteilung. Auch auf der Aggregatebene muss die Möglichkeit in Betracht gezogen werden, dass die durchschnittliche Leistung zwischen verschiedenen Organisationseinheiten unterschiedlich genau abgebildet und zum Besseren oder Schlechteren hin verzerrt wird. Aus den Erkenntnissen der Testtheorie und der Schulforschung lassen sich verschiedene Faktoren ableiten, die im Vergleich zwischen Fächern, Abschlussarten und/oder Hochschulen möglicherweise zu unterschiedlichen Durchschnittsnoten bei gleichen Leistungen führen.

Die fachspezifische Einbettung der Messung und Beurteilung von Kompetenzen lässt zudem erwarten, dass zusätzliche Einflüsse auf die Notengebung bestehen, die entsprechende Differenzen im Notenniveau zwischen Fächern bzw. Studiengängen begünstigen. Die fachspezifischen Prüfungsbedingungen lassen sich zum Teil über einen langen Zeitraum zurückverfolgen, weshalb mit langfristig stabilen fachspezifischen Mustern der Notengebung zu rechnen ist. Es gilt zu überprüfen, ob unterschiedlich geartete Distinktionskriterien (epistemologische, wissenschaftsorganisatorische und kulturelle Merkmale) zwischen den Fächern womöglich über unterschiedliche Mechanismen an unterschiedlichen Stellen des Notengebungsprozesses Einfluss nehmen und zu Differenzen im Notenniveau führen.

Die Existenz fachspezifischer Prüfungsbedingungen bedeutet jedoch nicht, dass auch zwangsläufig ein facheinheitliches Notenniveau an allen Hochschulen zu erwarten ist. Auch innerhalb von Fächern existieren unterschiedliche Prüfungsbedingungen an verschiedenen Hochschulen, die hochschulspezifische Notenniveaus begünstigen. Diese Abweichungen vom durchschnittlichen Fachniveau der Noten sind jedoch im Gegensatz zu den fachspezifischen Notenniveaus aufgrund hochschulinterner Veränderungsprozesse vermutlich eher anfällig für Brüche im Zeitverlauf.

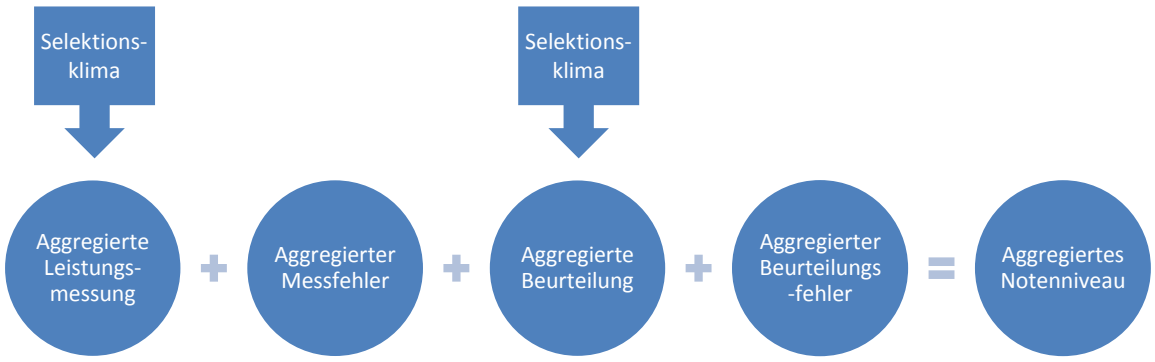
Die Komplexität der Analyse vervollständigend, muss neben der Berücksichtigung leistungsexterner Einflüsse auf die Notengebung als Ursache für Unterschiede im Notenniveau zwischen Fächern bzw. Studiengängen und innerhalb dieser stets beachtet werden, dass Noten trotz aller Verzerrungspotentiale immer noch Leistungsindikatoren darstellen und Unterschiede im Notenniveau zwischen verschiedenen Prüflingsgruppen prinzipiell auch immer durch unterschiedliche Leistungsniveaus der Gruppen zustande kommen können. Tabelle 1 fasst die möglichen Ursachen für unterschiedliche Notenniveaus in Abhängigkeit ihrer Anschlussstellen im Notengebungsprozess in zentrale Kategorien zusammen. Sowohl die Leistungsmessung als auch die Beurteilung enthalten, wie erläutert, prozessinhärente Potentiale der Produktion systematischer Unterschiede im Notenniveau. Je nach fach-, abschluss- und/oder hochschulspezifischer Ausgestaltung der Prüfungsbedingungen können diese

Potentiale in unterschiedlichem Maße zum Tragen kommen. Zusätzlich besteht die Möglichkeit, dass fach-, abschluss- und/oder hochschulspezifische Einflüsse zu unterschiedlich starken Selektionsneigungen der Prüfenden führen. Diese spezifischen Selektionsklimata üben möglicherweise Einfluss auf die Beurteilung von Leistungen aus, beeinflussen möglicherweise auch schon vor der Prüfung den Schwierigkeitsgrad der Aufgaben (Abb.2).

Tabelle 1: Mögliche Ursachen für unterschiedliche Notenniveaus (Kategorien)

Leistungskonforme Ursachen - Unterschiedliches Leistungsvermögen durch unterschiedliche:	Leistungsunabhängige Ursachen - Unterschiedliche Bewertungsstandards durch unterschiedliche:
Leistungsmessung - unterschiedliche Lehrqualität (Leistungsentwicklung) - unterschiedliche Zusammensetzung der Studierenden (Selbstselektion)	**Leistungsmessung** - unterschiedliche formale Prüfungsbedingungen - unterschiedliche Prüfungsverfahren - unterschiedliche Standardisierungsgrade
	Beurteilung - unterschiedliche Bezugsnormenorientierung - unterschiedliche Häufung von Wahrnehmungsfehlern
	Messung und Beurteilung - Selektionsklima - unterschiedliche Forschungsintensität - unterschiedlich hohe Prüfungsbelastung - unterschiedliche Rahmenbedingungen für Lehre - unterschiedliche Zusammensetzung der Lehrenden (z.B. Geschlecht, Alter, politische/pädagogische Einstellungen) - unterschiedlich gute Arbeitsmarktchancen der Prüflinge - unterschiedliche Finanzierungsstrukturen - unterschiedliche wissenschaftliche Ausrichtung (rein vs. angewandt) - unterschiedliche Rolle von Ideologien

Abbildung 2: Makroanalytische Darstellung des Notengebungsprozesses



6. Die Notengebung an Hochschulen – Empirische Befunde

Die Notengebung an deutschen Hochschulen ist ein im internationalen Vergleich kaum erforschtes Gebiet. Der Fokus der deutschen Bildungsforschung liegt generell, nicht erst seit internationalen Vergleichsstudien wie IGLU und PISA, auf dem Bereich der Schulforschung. Und auch innerhalb der deutschen Hochschulforschung findet die Notengebung im tertiären Bildungsabschnitt im Schatten von Themen wie Governance und Qualitätsmanagement kaum Beachtung, und wenn doch, dann meist in Form von Ratgebern zur Prüfungsvorbereitung und -durchführung.

So ist es kaum verwunderlich, dass im Vergleich zu zahlreichen Studien aus dem englischsprachigen Raum (v.a. USA, aber auch UK und Canada) zur Entwicklung von College- und Universitätsnoten nur eine geringe Anzahl deutscher Studien existieren, die sich vorrangig (größtenteils deskriptiv) mit diesem Thema beschäftigen. Neben einzelnen studiengangspezifischen Untersuchungen (z.B. Brinkmann (1967) und Ottwaska (1971) zu wirtschaftswissenschaftlichen Diplomprüfungen an der Universität zu Köln bzw. an der Universität Mannheim, Towfigh et al. (2014) zum ersten juristischen Staatsexamen, van den Bussche et al. (2006) zur ärztlichen Vorprüfung), konnten in einer umfassenden Recherche 12 Arbeiten, die sich der Vergabe von Hochschulnoten aus fachübergreifender Perspektive widmen, gefunden werden. Fünf davon weisen einen regional begrenzten Charakter auf: Eine Auswertung von Abschlussnoten an nordrhein-westfälischen Hochschulen durch von Dietrich (1984), eine an der Universität Marburg durch Hampe (1977; 1978) und zwei an der Universität des Saarlandes durch Apenburg et al. (1976; 1977a). Außerdem eine Analyse der Abschlussnoten mehrerer Studiengänge mit dem Abschluss des ersten Staatsexamens Lehramt an der Pädagogischen Hochschule Niedersachsen in Lüneburg von Ziegenspeck (1999). Die übrigen sieben Studien beruhen auf zwei Datenquellen. Sowohl Müller-Benedict und Tsarouha (2011) als auch der Wissenschaftsrat (2003; 2007; 2012) werten die amtliche Prüfungsstatistik aus, während die Arbeiten von Bitz (1989) und Maiworm (1989) beide Sonderauswertungen bzw. Erweiterungen einer hochschul- und fachübergreifenden Studie von Hitpass und Trosien (1987) darstellen[36].

Der im Folgenden zusammengefasste Stand der Forschung zur Notengebung an Hochschulen bezieht sich deshalb vor allem hinsichtlich der erklärenden Komponenten vorwiegend auf nicht-deutsche Verhältnisse, was eine Übertragbarkeit der Ergebnisse einschränkt. Bedingt durch den Mangel an Forschung im deutschen Hochschulsystem stellen die internationalen Forschungsergebnisse dennoch einen sinnvollen Ausgangspunkt für die Analyse der Hochschulnoten an deutschen Hochschulen dar, wobei die internationalen Unterschiede in den Bildungssystemen stets berücksichtigt werden müs-

[36] Im Ergebnisteil und im Fazit (Kapitel 8 und 9) finden sich außerdem Verweise auf Analyseergebnisse von Grözinger (2017) und McGrory (2017). Diese Ergebnisse entstanden ebenfalls im Rahmen des DFG-Forschungsprojekts „Die Notengebung an Hochschulen in Deutschland von den 1960er Jahren bis heute. Trends, Unterschiede, Ursachen.“, auf dem diese Arbeit beruht.

sen, sollten theoretische Erwägungen übertragen werden. Eine entsprechende Plausibilisierung der vorgestellten Erklärungsansätze folgt im Laufe der Arbeit.

6.1 Eine Frage der Perspektive: Querschnitt vs. Längsschnitt

Im Fokus von Analysen der Notengebung steht in der Regel das durchschnittliche Notenniveau, also das arithmetische Mittel der in einer bestimmten Untersuchungseinheit erzielten Noten, sowie gelegentlich deren Streuung. Notendurchschnitte können auf verschiedenen Untersuchungsebenen dargestellt werden, etwa auf Fachebene, auf Abschlussebene oder auf Hochschulebene, wobei Unterschiede und Gemeinsamkeiten zwischen Fächern, Abschlüssen und Hochschulen sichtbar werden. Vor allem in den USA werden die Noten nicht nur auf Fach- oder Studiengangebene, also über einzelne Hochschulen in einem Fach bzw. Studiengang, sondern häufig auch über verschiedene Fächer oder Studiengänge hinweg an einer Einrichtung zu einem Hochschuldurchschnitt aggregiert. Hier ist fraglich, inwiefern diese Werte noch einen vergleichenden Gehalt besitzen, da sich die einzelnen Hochschulen häufig deutlich in der Vielfalt und Anzahl der dort vertretenen Fächer bzw. Studiengänge (die spezifische Notenniveaus aufweisen) unterscheiden. Alternativ zur Betrachtung von Hochschulen als Aggregate der dort vertretenen Fächer und Abschlüsse, dürfte es sinnvoller sein, die Hochschulebene als Analyseebene unterhalb der Fach- bzw. Studiengangebene anzusiedeln und hochschulspezifische Notenvergaben innerhalb von Fächern oder Studiengängen zu analysieren.

In der Querschnittsperspektive können Notenniveaus für unterschiedliche Zeiträume bestimmt werden. So können zum Beispiel die Noten eines bestimmten Semesters oder eines bestimmten Jahres betrachtet werden oder aber Durchschnittswerte für mehrere Semester oder Jahre. Neben der Querschnittsperspektive besteht außerdem die Möglichkeit, Notenniveaus im Längsschnitt zu betrachten, also einen Vergleich einer oder mehrerer Untersuchungseinheiten zu einem bestimmten Zeitpunkt (oder mehreren Zeitpunkten) mit einem (oder mehreren) anderen Zeitpunkt(en) zu ziehen. Je nach Perspektive ergibt sich eine unterschiedliche Akzentuierung der Analyse: Die Querschnittsanalyse bringt Unterschiede und Gemeinsamkeiten im Notenniveau zwischen Untersuchungseinheiten hervor, die Längsschnittanalyse ermöglicht (vergleichende) Aussagen darüber, wie sich die Höhe dieses Notenniveaus im Zeitverlauf verändert und welche Form die Entwicklung des Notenniveaus annimmt. Sollten sich die erwarteten Unterschiede im Notenniveau zwischen Fächern bzw. Hochschulen nachweisen lassen, ist zu erwarten, dass auch im Zeitverlauf fachspezifische Entwicklungen auftreten - immer in Abhängigkeit der Entwicklung der für die Differenzen verantwortlichen Einflussfaktoren.

6.2 Das Notenniveau im Querschnitt

Die Querschnittsanalyse von Noten ist vor allem nützlich, um Unterschiede in der durchschnittlichen Höhe der Noten zwischen Fächern oder Studiengängen (an einer Hochschule oder über mehrere

Hochschulen gemittelt) sowie zwischen Hochschulen innerhalb eines Fachs oder Studiengangs zu identifizieren. Unterschiede im Notenniveau zu gegebenen Zeitpunkten zwischen Fächern und Hochschulen (sowohl über Fächer gemittelt als auch innerhalb von Fächern) werden in den USA schon seit den 1960er Jahren thematisiert. Relativ konstant wird seitdem davon berichtet, dass in den Erziehungswissenschaften (einschließlich der Lehrer*innenausbildung) die besten Noten vergeben werden, in den Geistes- und Sozialwissenschaften durchschnittlich bessere Noten als in den Naturwissenschaften (etwa Achen/Courant 2009; Goldman/Hewitt 1975; Koedel 2011; Weiss/Rasmussen 1960).

Fach- bzw. Fächergruppenunterschiede werden dabei in der US-amerikanischen Forschung vor allem an einzelnen Colleges und Universitäten (etwa Barth et al. 2009; Cheong 2000; Gamson 1967; Goldman/ Widawski 1976; Jewell et al. 2013; Johnson 2003; Prather et al. 1979 sowie Strenta/Elliot 1987 für Unterschiede zwischen einzelnen Kursen innerhalb der Betriebswirtschaftslehre) nachgewiesen, seltener hochschulübergreifend (wie bei Koedel 2011). Auch Unterschiede zwischen Hochschulen in ein und demselben Fach werden nur selten in der Querschnittsperspektive betrachtet (wiederum ebd. oder auch Connolly/Smith 1986 als Ausnahme). Zudem ist zu beachten, dass die US-amerikanischen Studien teils unterschiedliche Noten betrachten. Teilweise werden Abschlussnoten verwendet, teilweise Durchschnittsnoten während des Studiums und teilweise Noten einzelner Kurse. Diese uneinheitliche Handhabung wird dadurch begünstigt, dass sich die Gesamtabschlussnote in den USA schon länger aus einzelnen Modulprüfungen zusammensetzt, als dies in Deutschland (seit Einführung des Bachelor/Master-Systems) der Fall ist. Im Gegensatz zum deutschen Hochschulsystem ist in den USA nicht die Trennung nach gleichwertigen Abschlussarten sondern nach undergraduate und graduate Studierenden relevant, weshalb die hier getroffene Differenzierung zwischen Fach und Studiengang in den dortigen Studien keine Rolle spielt.

Fach- und hochschulspezifische Unterschiede im Notenniveau wurden nicht nur in den USA und Kanada (Anglin/Meng 2000; Eaton/Eswaran 2008), sondern auch in europäischen Ländern, etwa in Großbritannien (Bourner/Bourner 1985; Chapman 1994 und 1997; Connolly/Smith 1986; Yorke et al. 1996; Yorke 2008 und 2009) und Italien (Bagues et al. 2008; De Paola 2008) festgestellt. In Deutschland sind fächergruppenspezifische Differenzen in der Notenverteilung erstmals 1981 auf dem von der Freien Universität (FU) Berlin ausgerichteten Symposium „Diplomprüfungen im Widerstreit“ diskutiert worden. Im Rahmen dieser Veranstaltung wurden Berechnungen der FU Berlin präsentiert, nach denen Studierende der VWL und in Jura innerhalb eines Immatrikulationsjahrgangs weit seltener die Note „gut“ oder besser erreichten (20% bzw. 23%) als Studierende der Physik und in Medizin (82% bzw. 85%) (Klose/Lange 1981).

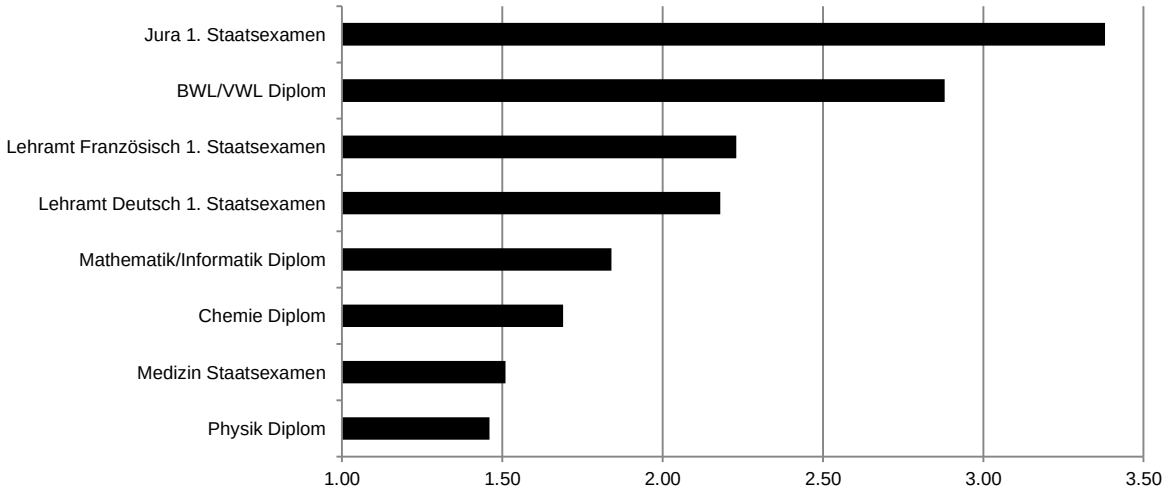
Wenige Jahre vor der öffentlichen Diskussion der Benotungspraxis an der FU Berlin präsentiert Hampe (1977 und 1978) die über den Zeitraum von Sommersemester 1964 bis Sommersemester 1968

gemittelten Abschlussnoten der Marburger Examenskandidaten für die Studiengänge VWL, Jura (1977), Medizin und Zahnmedizin (1978). Die (auf männliche Kandidaten beschränkte) Mittelung der Noten ergibt die schlechtesten Durchschnittswerte für Juristen ($\bar{x}$=3.6), Volkswirte schneiden nur unwesentlich besser ab ($\bar{x}$=3.5). Zahnmediziner ($\bar{x}$=2.5) und Mediziner ($\bar{x}$=2.0) hingegen erhalten vergleichsweise gute Noten.

Auch Apenburg et al. veröffentlichen bereits 1976 die Ergebnisse von Zwischen- und Abschlussprüfungen des Zeitraums Wintersemester 1972/73 bis Wintersemester 1974/75 bzw. bis Sommersemester 1975 (nur Medizin sowie Lehramt Deutsch und Französisch) an der Universität des Saarlandes (Apenburg et al. 1976). Ein Jahr später ergänzen sie die Datenbasis um das Sommersemester 1976 (Apenburg et al. 1977a). Berücksichtigt sind in dieser Erhebung die Diplomstudiengänge Betriebswirtschaftslehre/Volkswirtschaftslehre, Mathematik/Informatik (jeweils zusammengefasst), Chemie und Physik, die ersten Staatsexamina in Jura und Medizin, sowie die Gymnasial-Lehramtsstudiengänge Deutsch und Französisch.

Die Notenverteilungen und die auf diesen beruhenden Durchschnittsbildungen zeigen deutliche Unterschiede im Notenniveau zwischen den Studiengängen auf (Abb.3, auch Apenburg et al. 1977a:45). Die schlechtesten Noten, über die berücksichtigten Prüfungsperioden gemittelt, werden im ersten juristischen Staatsexamen vergeben. Etwas besser fallen die Noten in BWL und VWL aus, über die beiden Lehramtsstudiengänge und Mathematik/Informatik hin zu Chemie und Medizin werden die Noten besser. In Physik wurden im Zeitraum zwischen 1972 und 1975 die besten Noten an der Universität des Saarlandes vergeben. Die Autor*innen stellen zudem in Medizin, Chemie, Jura, BWL/VWL und Physik mittelstarke positive Korrelationen zwischen der Notenhöhe und der Anzahl der Fachsemester fest, die darauf hinweisen, dass die Noten bei zunehmender Studiendauer schlechter werden. In Mathematik/Informatik und den beiden Lehramtsstudiengängen zeigt sich nur ein geringer Zusammenhang (Apenburg et al. 1976).

Abbildung 3: Durchschnittliche Abschlussnoten an der Universität des Saarlandes von WiSe 1972/73 bis max. SoSe 1975



Quelle: Apenburg et al. 1976, eigene Darstellung

Neben den Fachunterschieden lassen die Daten keine weiteren systematischen Differenzen erkennen. Zwar befinden sich bei einer Aufteilung in vier besser und vier schlechter bewertende Studiengänge, drei der vier Staatsprüfungen in der Gruppe mit den schlechteren und drei der vier Diplomprüfungen in der Gruppe mit den besseren Noten. Allerdings stellt mit dem medizinischen Staatsexamen eine staatliche Prüfung die zweitbesten Ergebnisse, genauso wie die wirtschaftswissenschaftlichen Studiengänge als Diplomprüfungen die zweitschlechtesten Ergebnisse produzieren. Die Daten deuten demnach zwar Abschlussunterschiede an, reichen aber nicht aus, um systematische Unterschiede zwischen Abschlussarten im Notenniveau zu belegen. Dass keine Studiengänge mit dem Abschluss Magister in der Erhebung enthalten sind, erschwert entsprechende Interpretationen ebenso wie das Fehlen von gesellschafts- und/oder geisteswissenschaftlichen Studiengängen einen Vergleich zwischen Fächergruppen verhindert.

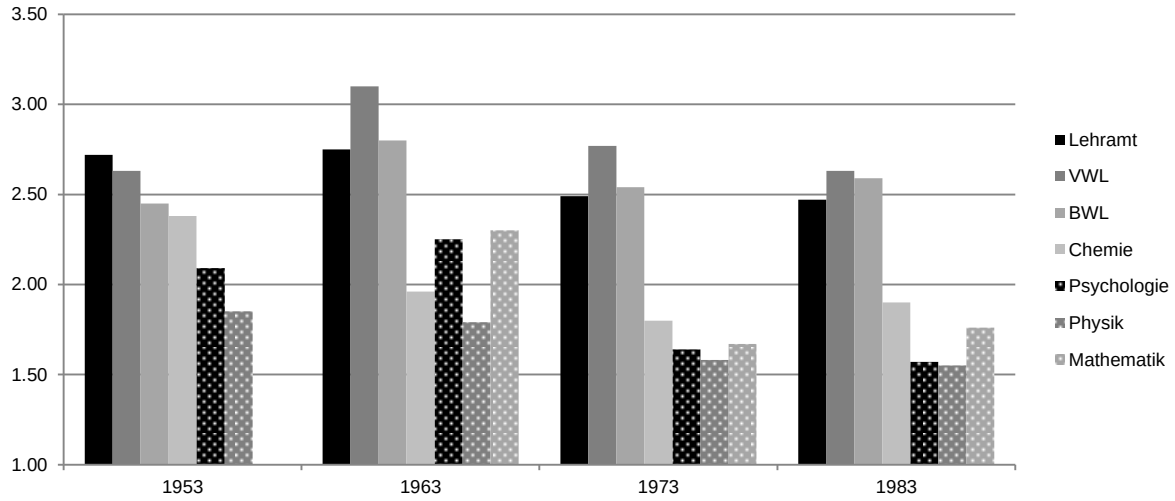
Einen solchen Fächervergleich ermöglicht von Dietrich (1984) in einer Auswertung der Abschlussnoten der Examensjahrgänge 1980 und 1982 an nordrhein-westfälischen Hochschulen. Aus seinen Daten geht hervor, dass in den Natur- und Geisteswissenschaften relativ betrachtet deutlich häufiger „gut“ oder „sehr gut“ als Prädikat vergeben wurden als in den Ingenieurswissenschaften, in denen wiederum ein größerer Anteil an Absolvent*innen eines dieser beiden Prädikate erhielt als in den Wirtschafts- und Sozialwissenschaften (von Dietrich 1984).

Hitpass und Trosien präsentieren 1987 die Abschlussnoten von 10 Fächern sowie des Studiengangs Lehramt für Sekundarstufe/Gymnasium (1. Staatsexamen), die sie stichprobenartig an mehreren Universitäten bundesweit zu vier Zeitpunkten erhoben haben[37]. Dabei variieren die einbezogenen Hochschulen zu den einzelnen Erhebungszeitpunkten 1953, 1963, 1973 und 1983 teilweise. Von den sechs der 10 Diplomstudiengänge, deren Ergebnisse umfassend dargestellt werden sowie den zusammengefassten Lehramtsstudiengängen (Biologie, Physik, Mathematik, Deutsch, Englisch und Geschichte) stellen sich die Physiknoten zu allen vier Erhebungszeitpunkten als beste Durchschnittsnoten dar. An drei Erhebungszeitpunkten werden sie gefolgt von den Noten in Psychologie, die besser sind als in Mathematik und Chemie (wobei Chemie zu einem Zeitpunkt besser als Psychologie ist). Lehramtsabsolvent*innen folgen zu drei der vier Messzeitpunkte auf Mathematik und Chemie, sind zu einem Zeitpunkt sogar die schlechtesten. Die Noten in BWL und VWL sind deutlich schlechter als in den anderen Studiengängen, wobei BWL-Absolvent*innen noch leicht besser abschneiden als ihre Kommiliton*innen in VWL. Lehramtsstudierende erzielen neben den Prüflingen in den beiden wirtschaftswissenschaftlichen Studiengängen die schlechtesten Noten (Hitpass/Trosien 1987). Innerhalb der Lehramtsstudiengänge lassen sich mit Englisch und Biologie ab 1963 zwei zeitlich konstant zu

[37] Für vier der 10 Diplomstudiengänge konnten Hitpass und Trosien nur unvollständige Informationen sammeln, weshalb sie diese vier auch nur in geringem Umfang behandeln.

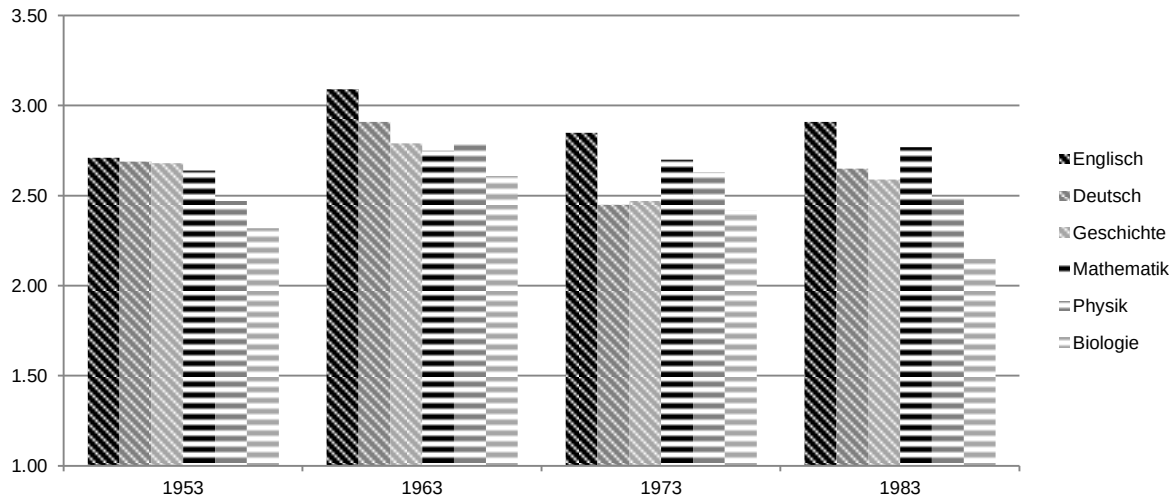
bleiben scheinende Extrempole finden, zwischen denen die vier anderen Studiengänge in jeweils unterschiedlichem Verhältnis zueinander liegen[38].

Abbildung 4: Durchschnittliche Abschlussnoten in sechs Diplomstudiengängen und im Lehramt (1.Staatsexamen)



Quelle: Hitpass/Trosien 1987, eigene Darstellung

Abbildung 5: Durchschnittliche Abschlussnoten in sechs Fächern mit Abschluss Lehramt Gymnasium (1. Staatsexamen)



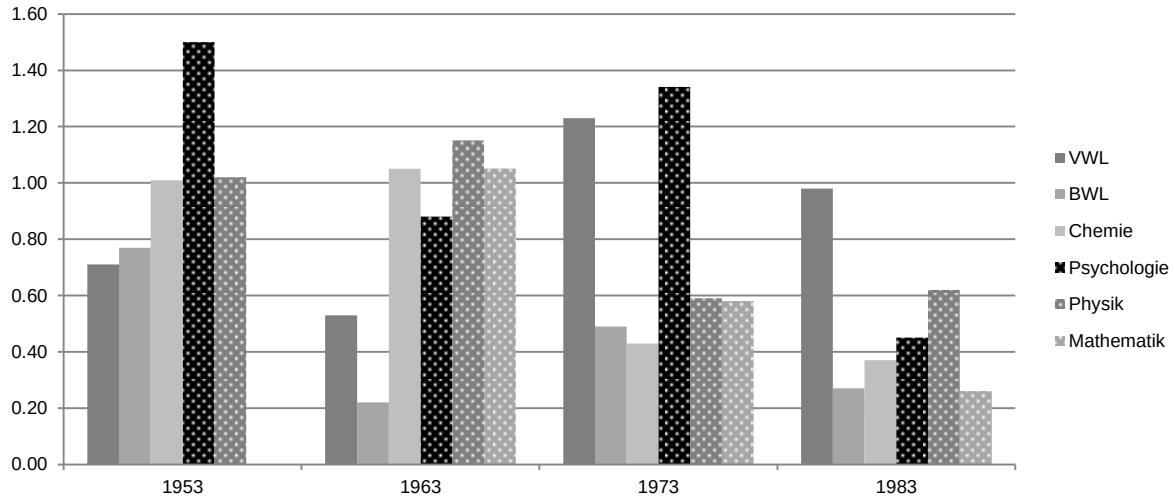
Quelle: Hitpass/Trosien 1987, eigene Berechnungen

Neben der erstmaligen Betrachtung der Notenentwicklung im Zeitverlauf beinhaltet die Studie von Hitpass und Trosien auch erstmals Daten, die einen systematischen Vergleich des Notenniveaus zwischen Hochschulen in ein und demselben Studiengang ermöglichen. Es zeigen sich in allen Studiengängen zu allen Messzeitpunkten deutliche, allerdings keineswegs konstante Spannweiten zwischen der Universität mit dem besten und der mit dem schlechtesten Notendurchschnitt. Über die vier Datenpunkte zeigt sich keine umfassende Systematik, es lässt sich lediglich ablesen, dass die Spannweiten in BWL und Chemie stets geringer ausfallen als in Physik. Bei der Betrachtung der einzelnen

[38] Siehe zu Unterschieden in den Durchschnittsnoten zwischen einzelnen Lehramtsfächern auch Ziegenspeck (1999): Dort beläuft sich die Spannweite gemittelt über den Erhebungszeitraum von 1965/66 bis 1972 auf R=0.7 für das Lehramt an Volksschulen.

Hochschulmittelwerte ist allerdings zu beachten, dass die Fallzahlen, die die Stichprobe von Hitpass und Trosien aufweist, für einige Hochschulen sehr gering ausfallen. Die Anzahl der Hochschulen, die in der Stichprobe enthalten sind, variiert zudem von Studiengang zu Studiengang[39].

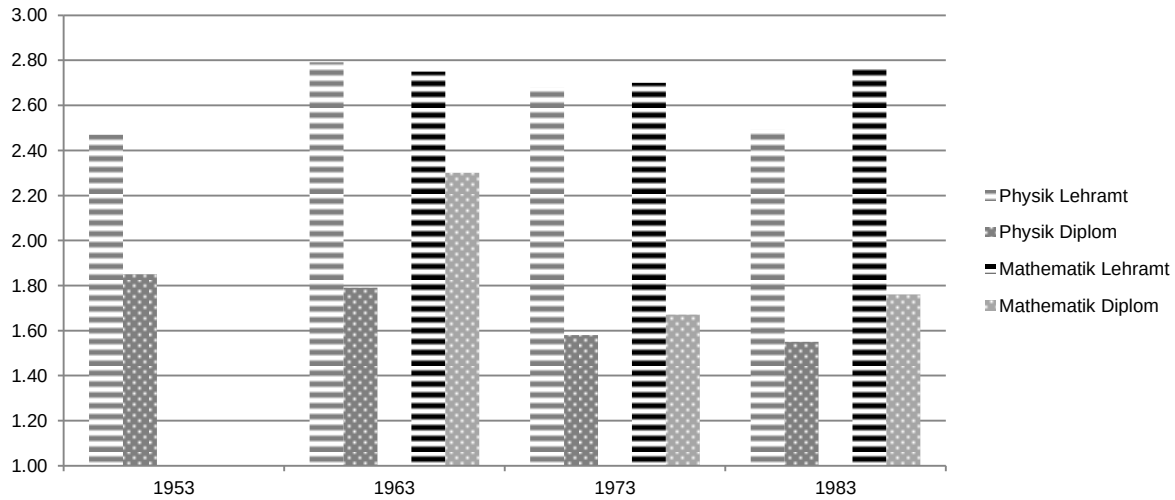
Abbildung 6: Spannweiten zwischen den einzelnen Hochschulen innerhalb der Diplomstudiengänge



Quelle: Hitpass/Trosien 1987, eigene Berechnungen

Doch nicht nur Fach- und Hochschulunterschiede werden in den Daten von Hitpass und Trosien deutlich. Die Autoren heben außerdem hervor, dass Abschlussnoten in der ersten Staatsprüfung der Lehramtsstudiengänge (hier Mathematik und Physik) deutlich schlechter ausfallen als in den entsprechenden Diplomstudiengängen, es also auch Notendifferenzen zwischen unterschiedlichen Abschlussarten gibt.

Abbildung 7: Durchschnittliche Abschlussnoten in Physik und Mathematik nach Abschluss

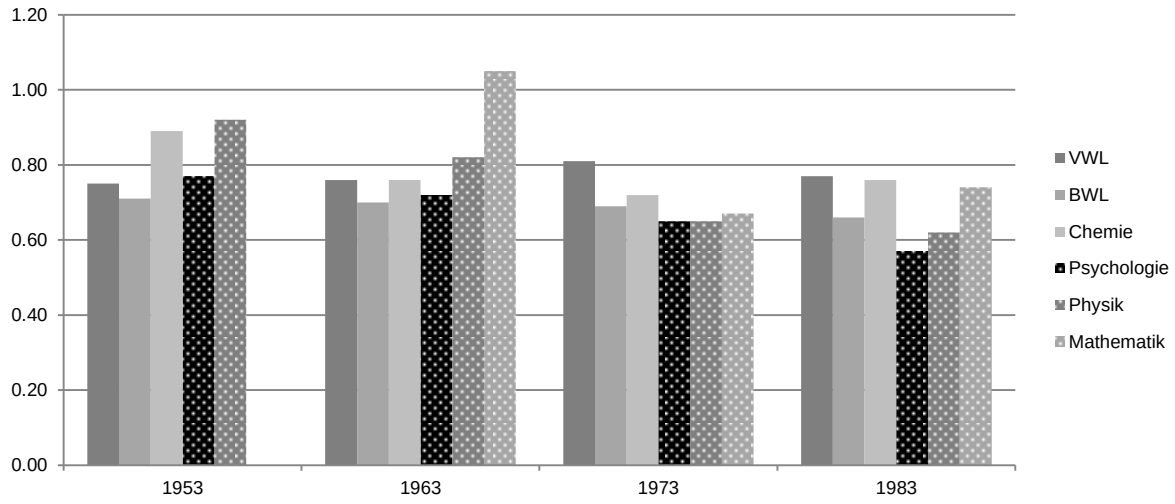


Quelle: Hitpass/Trosien 1987, eigene Berechnungen (nur Hochschulen mit Angaben in beiden Abschlussarten)

[39] Hitpass und Trosien präsentieren ihre Daten getrennt für die Hochschulen, die an allen vier Messzeitpunkten in der Stichprobe enthalten sind und für alle Hochschulen, für die sie Daten erhalten haben, dar. Die umfassendere Darstellung widmen sie ersterer Auswahl der gleich bleibenden Hochschulen, welche im Lehramt n=8, in Mathematik n=5, in Physik n=11, in Chemie n=3, in VWL n=4, in BWL n=2 und in Psychologie n=9 Hochschulen umfasst. Alle im Folgenden von Hitpass/Trosien entnommenen Daten beziehen sich auf diese an allen vier Messzeitpunkten gleichbleibenden Hochschulen.

Unterschiede lassen sich jedoch nicht nur im Notenniveau finden, also zwischen den gemittelten Notenwerten, sondern auch in der Streuung der Noten. Hitpass und Trosien führen neben den Fallzahlen und den Mittelwerten der einzelnen Hochschulen auch die jeweiligen Standardabweichungen der Notenmittelwerte auf. Aus diesen Parametern lassen sich die Standardabweichungen der über die Hochschulen auf Studiengangebene gemittelten Noten berechnen. Es fällt auf, dass die gemeinsamen Standardabweichungen deutlich geringere Unterschiede zwischen den fachlich abgrenzbaren Studiengängen aufweisen als die Durchschnittsnoten. Lediglich für 1963 zeigt sich ein nennenswerter Unterschied zwischen Mathematik und den Studiengängen der übrigen Fächer. Im Verhältnis einzelner Studiengänge zueinander lässt sich erkennen, dass die Streuung der Noten in VWL stets größer ausfällt als in BWL, in Chemie stets größer als in Psychologie und BWL. Ein systematisches Verhältnis, das sich zu allen vier Messzeitpunkten äußert, ist jedoch nicht auszumachen. Hinsichtlich der Streuung zeigen sich wie auch wie auch für die Mittelwerte der Noten deutliche Unterschiede zwischen den einzelnen Hochschulen innerhalb der Studiengänge. Da die Fallzahlen allerdings, wie bereits erwähnt, für einige Hochschulen sehr gering ausfallen, sind die Werte der einzelnen Hochschulen vermutlich wenig aussagekräftig.

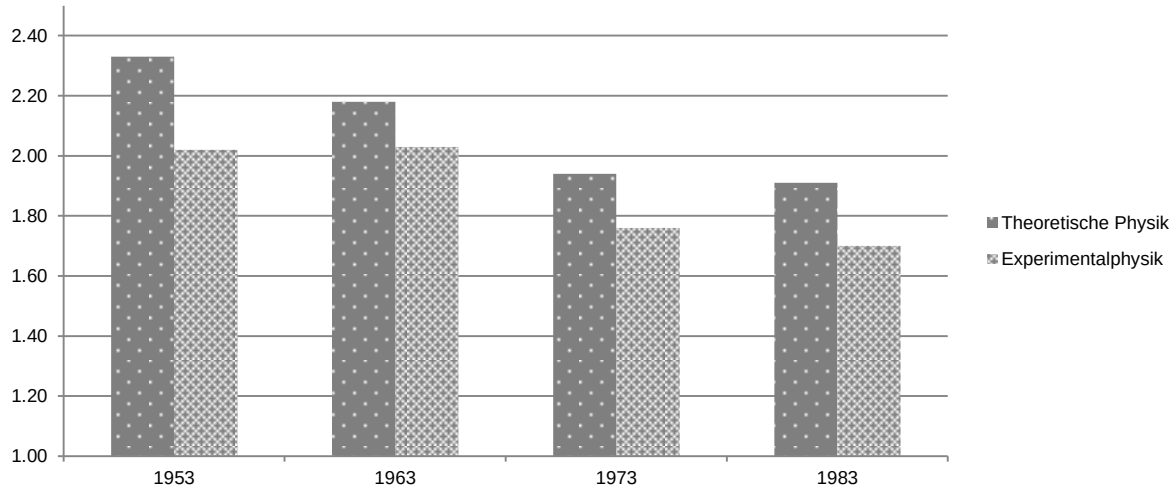
Abbildung 8: Gemeinsame Standardabweichung der Hochschulen auf Fachebene



Quelle: Hitpass/Trosien 1987, eigene Berechnungen

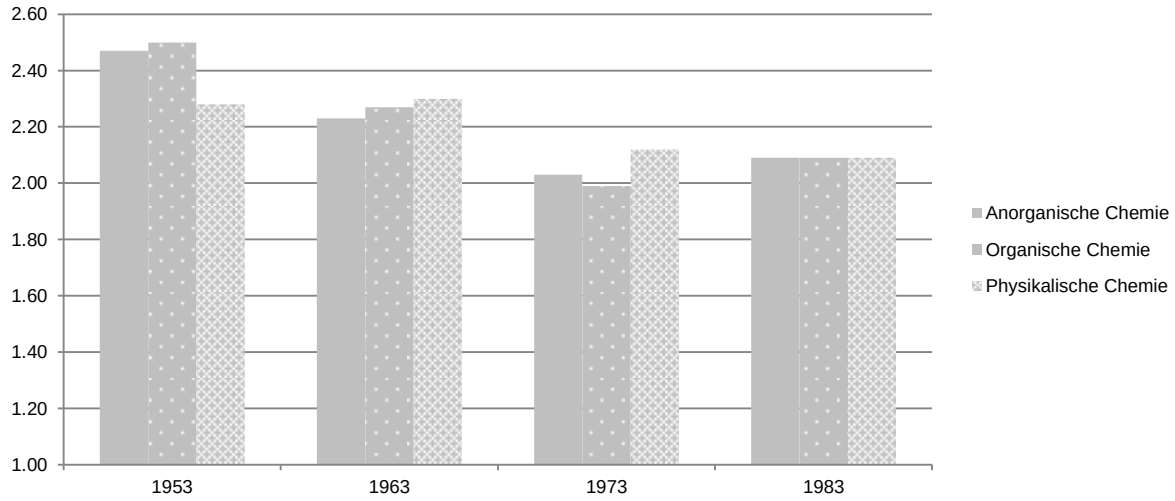
Maiworm (1987) führt anhand der von Hitpass und Trosien erhobenen Daten einen Vergleich der hessischen Hochschulnoten mit den Noten auf Bundesebene durch. Seine Berechnungen enthalten dabei unter anderem differenzierte Angaben zu den Noten in den einzelnen Teilgebieten der Fächer Chemie, Mathematik und Physik. Aus ihnen wird deutlich, dass sich die Unterschiede zwischen den Hochschulen auch innerhalb der Spezialgebiete widerspiegeln, zwischen diesen selbst allerdings weitaus geringfügigere Differenzen zu finden sind (siehe auch Bitz 1989 zu einem detaillierten Vergleich der Diplomprüfungen in Physik und Psychologie). Am ehesten zeichnen sich Unterschiede im Notenniveau zwischen den Teilgebieten noch in der Physik ab, auch hier erreichen sie aber maximal eine Spannweite von R=0.31 in 1953.

Abbildung 9: Durchschnittliche Abschlussnoten in Physik nach Teilgebiet



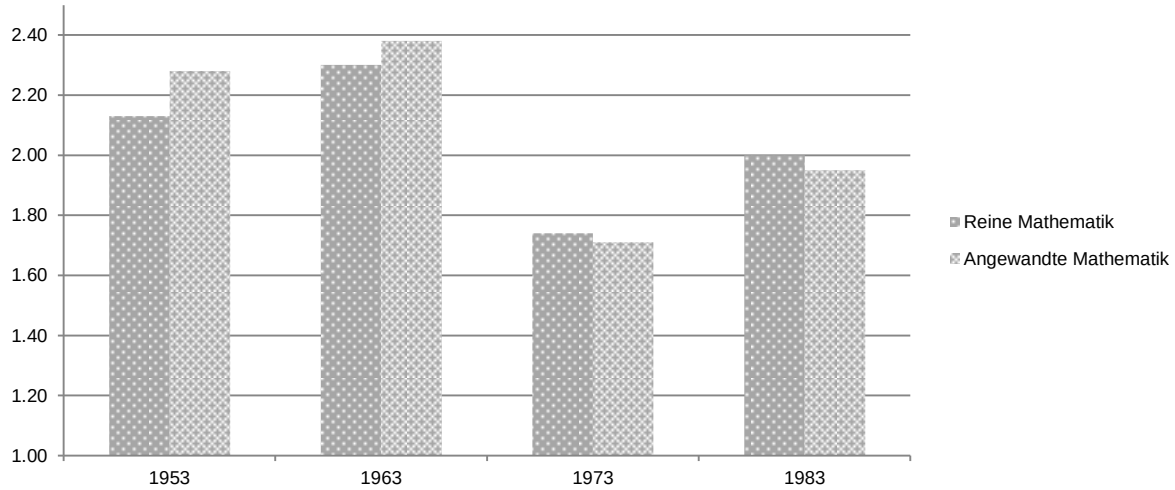
Quelle: Maiworm 1987, eigene Darstellung

Abbildung 10: Durchschnittliche Abschlussnoten in Chemie nach Teilgebiet



Quelle: Maiworm 1987, eigene Darstellung

Abbildung 11: Durchschnittliche Abschlussnoten in Mathematik nach Teilgebiet



Quelle: Maiworm 1987, eigene Darstellung

Die nächste wissenschaftliche Aufarbeitung von Prüfungsnoten an Hochschulen erfolgte 2003 durch den Wissenschaftsrat. Dieser veröffentlichte in den Jahren 2003, 2007 und 2012 deskriptive Auswer-

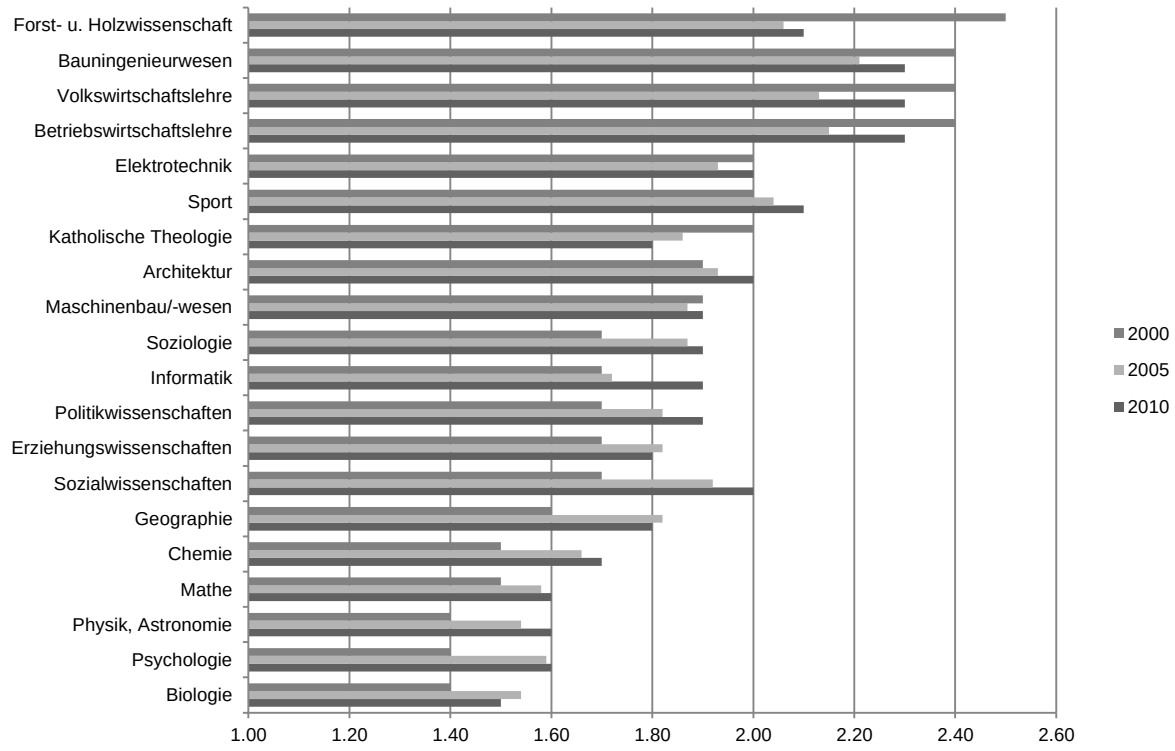
tungen der amtlichen Prüfungsstatistik, die frühestens ab 1993 die Prüfungsnoten aller Absolvent*innen bundesdeutscher Hochschulen enthält. Die Analysen beziehen sich im ersten Bericht auf die Examensnoten aus den Jahren 1996, 1998 und 2000, im zweiten Bericht auf die Noten des Jahres 2005 und im dritten Bericht auf die Noten aus dem Prüfungsjahr 2010. In der ersten Analyse aus 2003 weist der Wissenschaftsrat auf eine eingeschränkte Ausnutzung der Notenskala, die sich meistens im Bereich der besseren Noten bewegt, und auf deutliche Fachunterschiede hin:

> „Die Einzelergebnisse zeigen, dass in bestimmten Studiengängen für die Leistungsbewertung das Notenspektrum nicht ausgeschöpft wurde, verbunden mit einer geringen Notendifferenzierung. Außerdem wurden in einigen Fächern besonders gute Noten auffallend häufig vergeben. Die Durchschnittsnoten variierten von 1,3 in Biologie bis 3,3 in Rechtswissenschaften". (Wissenschaftsrat 2003:28)

Dass in den meisten universitären Studiengängen vor allem gute und sehr gute Noten vergeben werden, verdeutlicht der Umstand, dass 1996 von den ausgewiesenen 32 Diplom- und 18 Magisterstudiengängen an Universitäten nur sechs einen Notendurchschnitt schlechter als $\bar{x}$=2.0 aufweisen (1998 sind es sieben, 2000 sind es fünf Studiengänge), drei davon sind wirtschaftswissenschaftliche Studiengänge. In Kontrast dazu ist die Durchschnittsnote in fünf (1998 ebenfalls in fünf, 2000 in vier) von sechs Studiengängen mit Abschluss Staatsexamen schlechter als $\bar{x}$=2.0, nur in einem (2000 in zwei) $\bar{x}$=2.0 oder besser (Wissenschaftsrat 2003), was die bereits bei Hitpass und Trosien aufgezeigten Unterschiede im Notenniveau zwischen Abschlussarten bestätigt. Diese Unterschiede sowie die im Bericht angesprochene Varianz zwischen den Notendurchschnitten der Fächer lässt sich anhand einer Darstellung der Mittelwerte für einige Studiengänge für den Zeitraum 2000 bis 2010 nachzeichnen. Es zeigt sich hier, dass vor allem zwischen Diplomstudiengängen und zwischen Studiengängen mit Abschluss Staatsexamen große Unterschiede im Notenniveau existieren, während die Spannweiten im Magister und in Lehramtsstudiengängen deutlich geringer ausfallen. Die besten Noten mit Abschluss Staatsexamen gibt es in der Zahnmedizin, die schlechtesten, auch über die Abschlüsse hinweg, in den Rechtswissenschaften. Die über die Abschlüsse hinweg besten Noten werden im Durchschnitt in Biologie, Psychologie und Physik (jeweils Diplom) vergeben. Die schlechtesten Diplomnoten gibt es in den Forstwissenschaften und im Bauingenieurswesen sowie in den wirtschaftswissenschaftlichen Studiengängen. Im Magister werden in Musik, in den Literatur- und Sprachwissenschaften sowie in Philosophie und Kunst die besten Noten vergeben, im Lehramt liegen nur Kunst und Musik etwas unter den Noten der übrigen Fächer. In Romanistik, Anglistik, in Bibliothekswissenschaften, Geographie, Politikwissenschaften und Erziehungswissenschaften schneiden die Studierenden im Vergleich zu ihren Kommiliton*innen im Magister schlechter ab. Im Vergleich einzelner Fächer mit der unterschiedlichen Abschlussmöglichkeit Diplom vs. Magister zeigen sich nur geringe Unterschiede: Im Studiengang Sozialwissenschaften und auch in Soziologie und Politikwissenschaften sind die Noten in beiden Abschlussarten ungefähr gleich gut. In Geographie und den Erziehungswissenschaf-

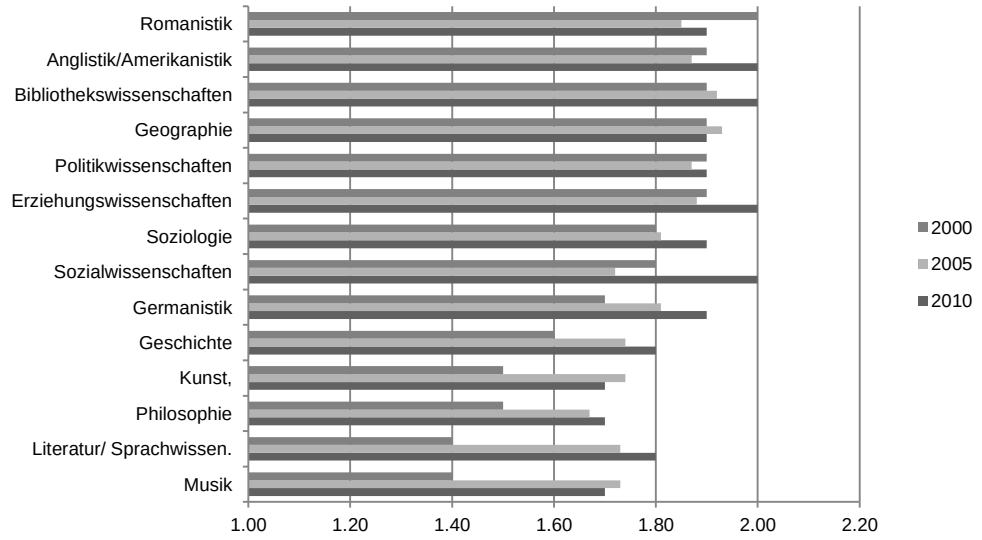
ten erhalten Magisterstudierende nur geringfügig schlechtere Noten als Diplomstudierende. Ein Vergleich der Lehramtsstudiengänge mit ihren Diplom- bzw. Magisteräquivalenten zeigt schlechtere Noten im Lehramt als im Diplom, aber nur unwesentliche Unterschiede gegenüber Magisterstudiengängen im gleichen Fach.

Abbildung 12: Abschlussnoten in ausgewählten Diplomstudiengängen



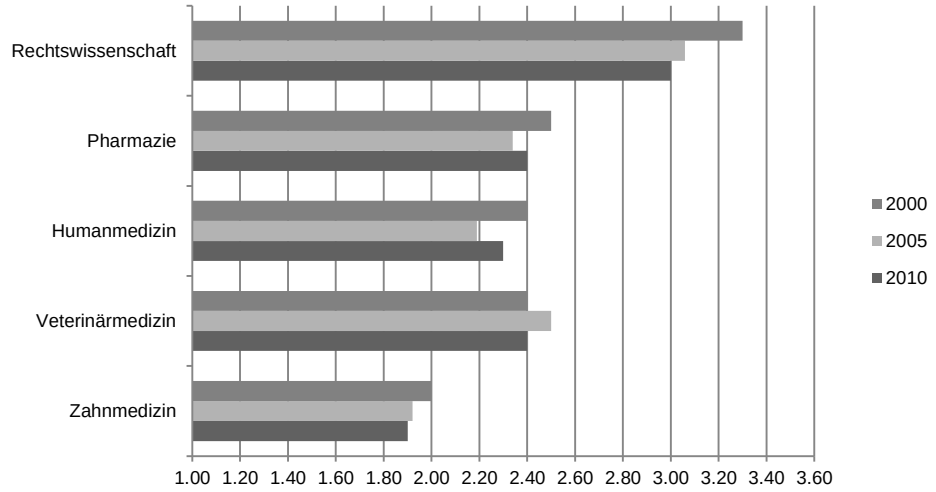
Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Darstellung

Abbildung 13: Abschlussnoten in ausgewählten Magisterstudiengängen



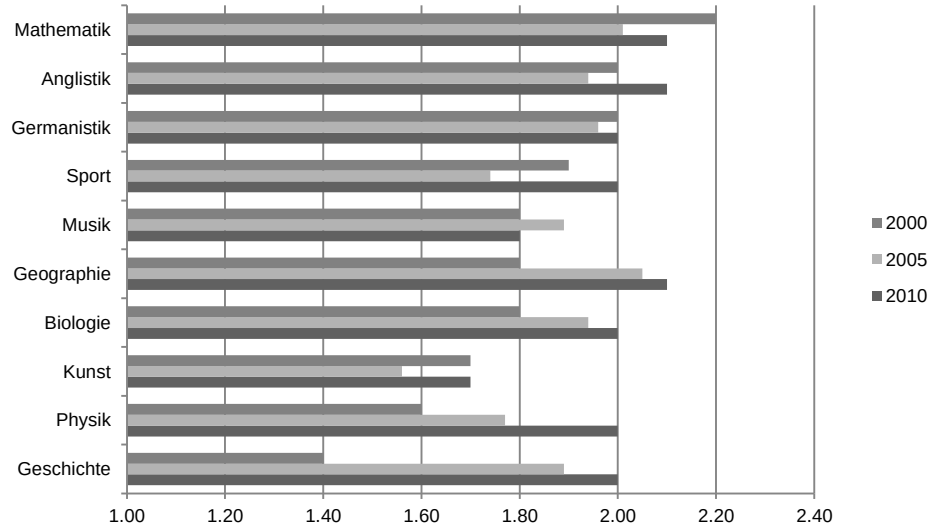
Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Darstellung

Abbildung 14: Abschlussnoten in ausgewählten Staatsexamensstudiengängen



Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Darstellung

Abbildung 15: Abschlussnoten in ausgewählten Lehramtsstudiengängen (Gymnasium)



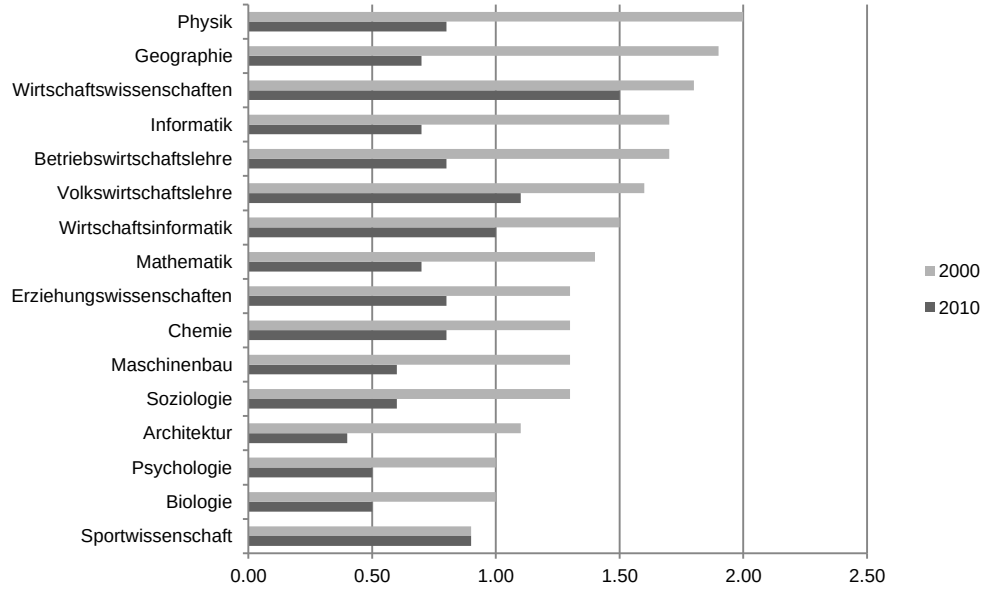
Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Darstellung

Auswertungen für das Prüfungsjahr 2000 zeigen zudem deutliche Unterschiede im Notenniveau innerhalb einzelner Studiengänge zwischen den Hochschulen, an denen sie angeboten werden. Die Spannweiten innerhalb der Studiengänge unterscheiden sich dabei in ihrem Ausmaß, sind am geringsten in den Diplomstudiengängen Sport, Psychologie und Biologie, am größten in den Magisterstudiengängen Politik, Kunst und Geschichte sowie im Diplom Physik (Abb.16 und 17). Die Existenz studiengang- wie hochschulspezifischer Notenniveaus bestätigt sich auch in den beiden folgenden Berichten des Wissenschaftsrats. In der Auswertung der Noten des Prüfungsjahrgangs 2005 wird explizit auf die bereits 2000 zu erkennenden Hochschulunterschiede hingewiesen:

„Die Ergebnisse der vorliegenden Auswertung für das Prüfungsjahr 2005 bestätigen im Großen und Ganzen die Befunde des Arbeitsberichts zu Prüfungsnoten für das Jahr 2000. Sie zeigen zum einen, dass die Prüfungsergebnisse zwischen verschiedenen Studiengängen kaum vergleichbar sind. Zum anderen weisen einige Studiengänge große Notendiskrepanzen zwischen den Hochschulen auf.“ (ebd. 2007:11)

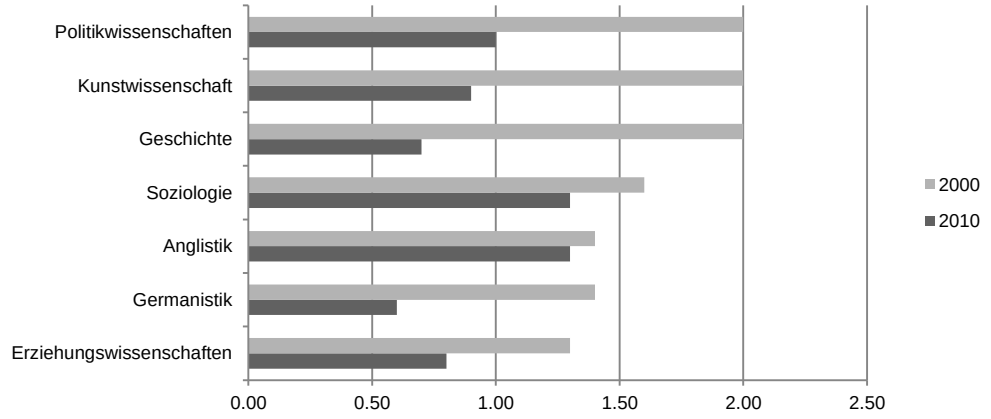
Auch für 2010 werden „sowohl zwischen den als auch innerhalb der Fachbereiche auffällige Spreizungen“ (ebd. 2012:7) hervorgehoben. Dabei zeigen die publizierten Daten, dass die Spannweiten zwischen den einzelnen Hochschulen in allen Fächern 2000 noch größer waren als 2010, auch 2010 jedoch, wie im Bericht erwähnt, deutliche Unterschiede in den Notenmittelwerten zu finden sind.

Abbildung 16: Spannweiten zwischen Hochschulen in ausgewählten Diplomstudiengängen 2000 und 2010



Quelle: Wissenschaftsrat,2003/2012, eigene Berechnungen

Abbildung 17: Spannweiten zwischen Hochschulen in ausgewählten Magisterstudiengängen 2000 und 2010

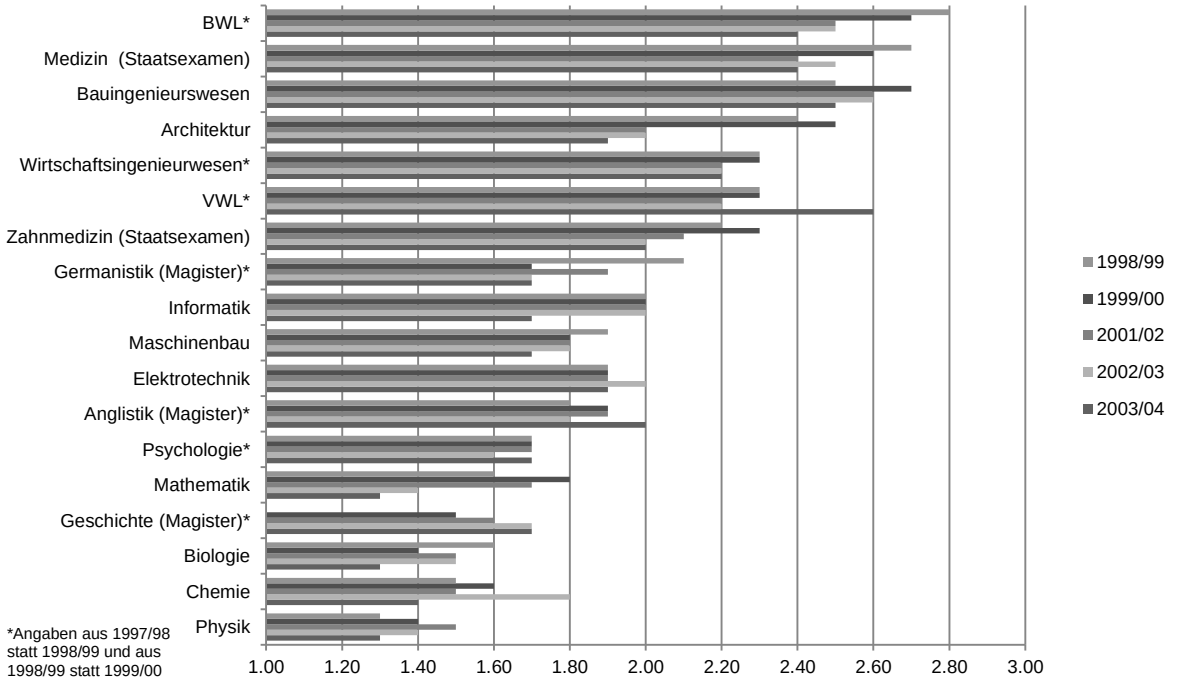


Quelle: Wissenschaftsrat,2003/2012, eigene Berechnungen

Doch nicht nur der Wissenschaftsrat hat das Thema Notengebung an Hochschulen in jüngerer Vergangenheit wieder aufgegriffen. Krempkow (2000-2005) präsentierte im Rahmen des „Studienführer Sachsen“ die nach Studiengängen differenzierten Abschlussnoten an den verschiedenen sächsischen Hochschulen und wies dabei explizit auf deutliche Studiengang- und Hochschulunterschiede hin. Auch hier zeigt sich mit wenigen Ausnahmen das bereits in den Daten des Wissenschaftsrats ersichtliche Muster: Die besten Noten gibt es in naturwissenschaftlichen, die schlechtesten in ingenieur-und wirtschaftswissenschaftlichen Diplomstudiengängen, während die Magisterstudiengänge zwischen den beiden Polen liegen (vgl. Abb.18 und 19 für die Universitäten Dresden und Leipzig).

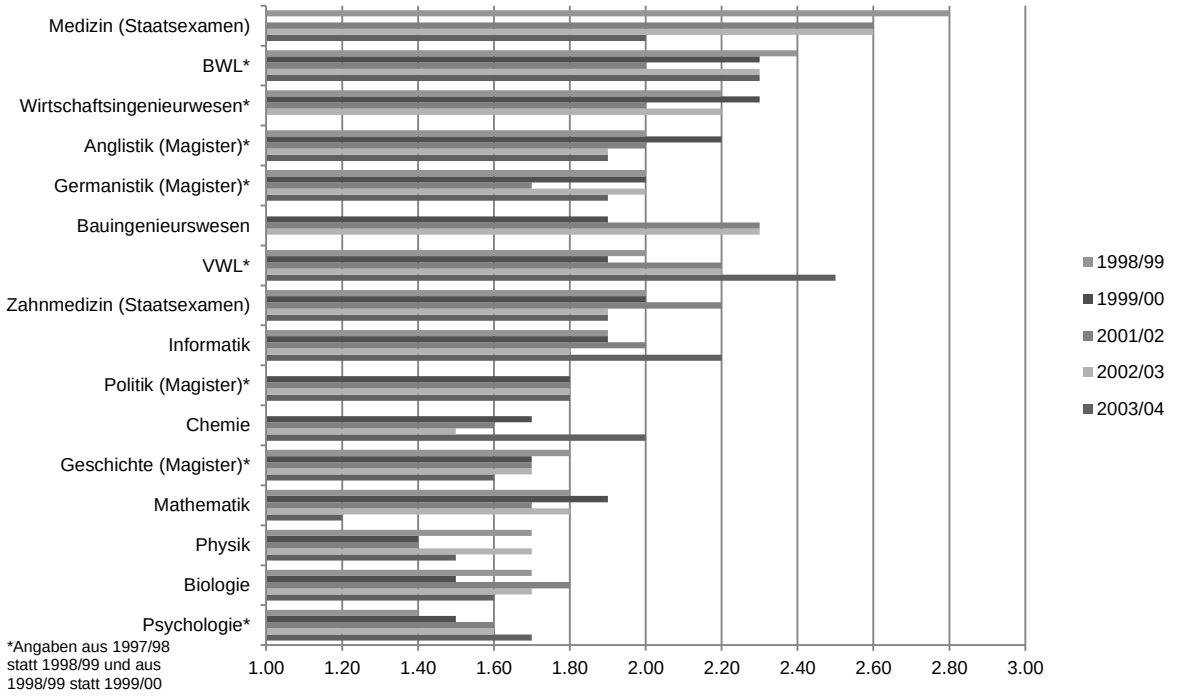
Müller-Benedict und Tsarouha unternahmen 2011 einen Versuch, die deutsche Notengebung systematischer als bisher geschehen zu beschreiben. Ebenfalls unter Rückgriff auf die amtliche Prüfungsstatistik können auch sie Unterschiede zwischen Abschlüssen, Fächern und Hochschulen im selben Fach nachweisen, wobei ihre Analysen im Gegensatz zu denen des Wissenschaftsrats auf den Noten aller Prüfungsjahrgänge seit dem Wintersemester 1995/1996 beruhen (Müller-Benedict/Tsarouha 2011). Die Autor*innen bestätigen anhand der amtlichen Daten die punktuellen Ergebnisse des Wissenschaftsrates: In Diplomstudiengängen erhalten Prüflinge bessere Noten als im Staatsexamen des gleichen Fachs, Prüflinge in Biologie, Physik und Psychologie erhalten die besten Noten, die in Jura die schlechtesten. Die Analysen von Müller-Benedict und Tsarouha zeigen erstmals, dass die zuvor punktuell festgestellten Unterschiede zwischen Abschlüssen, Fächern und Hochschulen auch über einen längeren Zeitraum hinweg stabil bleiben.

Abbildung 18: Abschlussnoten in ausgewählten Fächern an der TU Dresden



Quelle: Krempkow 2002/2003/2004/2005, eigene Darstellung

Abbildung 19: Abschlussnoten in ausgewählten Fächern an der Universität Leipzig



Quelle: Krempkow 2002/2003/2004/2005, eigene Darstellung

Die Zusammenfassung der empirischen Ergebnisse aus Querschnittsperspektive zeigt, dass Unterschiede in der Höhe des Notenniveaus zwischen Abschlussarten, Fächern und innerhalb der einzelnen Studiengänge zwischen Hochschulen unabhängig vom länderspezifischen Hochschulsystem existieren. FH2a (hochschulübergreifend zeitlich stabile Differenzen im Notenniveau zwischen fachlich abgegrenzten Studiengängen) ist damit bereits bestätigt, wobei noch unklar ist, wie sich die zeitliche Stabilität der Unterschiede genau darstellt[40]. In Deutschland zeigen sich im Gegensatz zu den USA übereinstimmend bessere Noten in den Naturwissenschaften als in den Geisteswissenschaften, die Rechtswissenschaften, ingenieur- und wirtschaftswissenschaftliche Studiengänge vergeben die schlechtesten Noten. Im Staatsexamen (mit und ohne Lehramt) werden schlechtere Noten vergeben als in Diplomstudiengängen des gleichen Fachs, zwischen Diplom- und Magisterstudiengängen des gleichen Fachs zeigen sich allerdings kaum Unterschiede, ebenso zwischen Lehramt und Magisterstudiengängen innerhalb eines Fachs. Deutliche Unterschiede zwischen den Fächern zeigen sich innerhalb aller Abschlüsse außer im Lehramt, wo die Notenniveaus der einzelnen Fächer nah beieinanderliegen.

Hinsichtlich der Streuung der Noten lässt sich zumindest aus den Daten von Hitpass/ Trosien keine Fachsystematik ablesen. Die Studien, die die Notengebung nicht nur auf Studiengang- sondern auch auf Hochschulebene erfassen, zeigen, dass auch innerhalb der Studiengänge Unterschiede im Notenniveau zwischen einzelnen Hochschulen existieren, wobei die Spannweiten stark variieren, aber kei-

[40] Dass Notenunterschiede im Querschnitt kein neues Phänomen sind, zeigen Müller-Benedict et al. (2008) anhand historischer Daten bereits für den Zeitraum von 1865-1941 für vier ausgewählte Karrieren.

ne Systematik dieser Differenzen erkennen lassen. FH4a (an einzelnen Hochschulen existieren signifikante, im Zeitverlauf stabile Abweichungen vom durchschnittlichen Notenniveau im jeweiligen Studiengang) kann durch diese Befunde erst einmal nur bedingt bestätigt werden, da aus den vorhandenen Daten nicht erkenntlich ist, inwiefern die Unterschiede zwischen den Hochschulen zeitliche Stabilität aufweisen. Schließlich ist festzuhalten, dass zwischen den Teilgebieten einzelner Fächer keine nennenswerten Differenzen im Notenniveau zu existieren scheinen.

Zur Erklärung der festgestellten Unterschiede zwischen Abschlüssen, Fächern und Hochschulen können, wie in Kapitel 2 beschrieben, grundsätzlich zwei Hauptdimensionen herangezogen werden: Leistungskonforme und leistungsunabhängige Faktoren. So ist es naheliegend, bessere Noten mit besserer Leistung zu erklären, schließlich sollen Noten Leistung abbilden. Dieser Erklärungslinie folgend, können Leistungsunterschiede zwischen Studierenden verschiedener Abschlüsse, Fächer und Hochschulen auf zwei Arten zustande kommen. Einerseits besteht die Möglichkeit, dass die Studierenden sich schon zu Beginn ihres Studiums in leistungsdeterminierenden Dispositionen unterscheiden, dass vergleichsweise begabtere Studierende bestimmte Studiengänge, Abschlussarten und Hochschulen präferieren, andere meiden (Chapman 1994). In der Literatur zur Notengebung an US-amerikanischen Hochschulen werden als Kriterien der Zusammensetzung, die möglicherweise die Leistung beeinflussen das Geschlecht, die soziale Herkunft, das Alter, die Studienerfahrung, die Ethnizität sowie der Status als Transferstudierende*r, Teilzeitstudierende*r und/oder Stipendiat*in genannt (etwa Kuh/Hu 1999; Lowe et al. 2008; Mathies/Webber 2009). Ein solcher Selektionsprozess müsste sich durch bestimmte Distinktionsmerkmale der Studiengänge und Hochschulen erklären lassen, die dafür verantwortlich sind, bestimmte Fähigkeitsklassen von Studierenden anzuziehen.
Andererseits kann sich die Leistung der Studierenden im Laufe ihres Studiums unterschiedlich stark entwickeln. Hierfür wären systematische Unterschiede in der Lehrqualität verantwortlich, die sich wiederum auf abschluss-, fach- oder hochschulspezifische Rahmenbedingungen der Lehre zurückführen lassen müssten, etwa auf unterschiedliche finanzielle oder personelle Ausstattungen (ebd.). Schließlich ist ein Zusammenwirken beider Mechanismen denkbar. So könnte sowohl bessere Lehrqualität einen Grund für einen Selektionsprozess darstellen als auch bessere Lehrqualität erst dort entstehen, wo die Studierenden die Lehre engagierter und befähigter aufnehmen.
Lassen sich Unterschiede im Notenniveau zwischen den betrachteten Untersuchungseinheiten nicht oder nicht vollständig durch Unterschiede in der Leistung der Studierenden erklären, sind zwangsläufig leistungsunabhängige Einflussfaktoren für die festgestellten Differenzen (mit-)verantwortlich. Sie könnten Unterschiede im Notenniveau, die unabhängig von möglichen Leistungsunterschieden zwischen Prüflingen bestehen, also unterschiedliche *Bewertungsstandards* zwischen Fächern, Hochschulen und Abschlüssen erklären.

Um herauszufinden, zu welchem Anteil Unterschiede im Notenniveau durch unterschiedliche Leistungsniveaus zu Studienbeginn entstehen, werden in der US-amerikanischen Forschung in der Regel die Ergebnisse von College- und Universitätseingangstests (in der Regel **S**cholastic **A**ssessment **T**est- bzw. **A**merican **C**ollege **T**esting-Scores) herangezogen. Die Unterschiede im Notenniveau zwischen Fächern oder Hochschulen werden dann in Relation zu den Unterschieden in den Ergebnissen dieser standardisierten Leistungstests gesetzt, um aufzuzeigen, welcher Anteil der Varianz in den Noten durch Unterschiede in den Fähigkeiten der Studierenden erklärt werden kann. Der restliche Anteil wird dann als leistungsunabhängig determiniert betrachtet. Da in Deutschland keine standardisierten Tests zur Erhebung studentischer Fähigkeiten durchgeführt werden, muss ein Selektionsprozess zu Beginn des Studiums auf andere Weise überprüft werden.

Müller-Benedict und Tsarouha lösen dieses Problem, indem sie die durchschnittlichen Abiturnoten von Studierenden unterschiedlicher Abschlussarten vergleichen. Da die Abiturnote den besten Einzelprädiktor für den Studienerfolg darstellt (Trapmann et al. 2007; zusammenfassend: Köller 2013) kann davon ausgegangen werden, dass sie einen aussagekräftigen Indikator der Eingangseignung von Studierenden darstellt. In studiengang- bzw. hochschulspezifischen Studien zu den Bedingungen des Studienerfolgs konnte dementsprechend mehrfach ein Zusammenhang zwischen Abiturgesamtnote und Prüfungsleistungen im Studium nachgewiesen werden (Brinkmann 1967; Erdel 2010; Giese et al. 2013; Towfigh et al. 2014). Müller-Benedict und Tsarouha können für den Vergleich nach Abschlüssen keine nennenswerten Unterschiede in der Abiturnote feststellen, was gegen eine Erklärung der abschlussspezifischen Notenniveaus durch Selektion der fähigeren Studierenden in Studiengänge mit bestimmten Abschlussarten spricht (Müller-Benedict/Tsarouha 2011).

Auch hinsichtlich der Differenzen im Notenniveau zwischen einzelnen Hochschulen im gleichen Fach zeigen Müller-Benedict und Tsarouha, dass eine Selektion nach unterschiedlicher Eingangseignung allein die Unterschiede nicht erklären kann. Sie vergleichen hierzu die Noten von Hochschulen in Studiengängen mit Numerus Clausus (NC), in denen eine Selbstelektion aufgrund der zentralen Studienplatzvergabe nur bedingt (durch die Berücksichtigung von Standortpräferenzen) möglich ist. Hier sollte sich die Eingangseignung aufgrund der Zuweisung zu einer Hochschule nicht systematisch zwischen den Hochschulen unterscheiden. Dennoch bestehen auch innerhalb von NC-Studiengängen deutliche Spannweiten zwischen den Hochschulen (ebd.). Einschränkend muss bei diesem Vorgehen jedoch darauf hingewiesen werden, dass auch in den NC- Studiengängen hochschulspezifische Unterschiede durch die Eingangseignung möglich sind, und zwar durch die unterschiedlichen bundeslandspezifischen NC-Werte. Van den Bussche et al. (2006) zeigen für die ärztliche Vorprüfung, dass mit höherem bundeslandspezifischem NC auch schlechtere Durchschnittsnoten einhergehen.

Inwiefern sich die Noten zwischen den Fächern aufgrund unterschiedlicher Begabung der Studierenden erklären lassen prüfen Müller-Benedict und Tsarouha zwar nicht dezidiert, sie präsentieren aller-

dings Notenniveaus aus NC- Studiengängen wie auch Nicht-NC- Studiengängen und auch den Wissenschaftsberichten lassen sich die entsprechenden Noten entnehmen. Sollte der Erklärungsansatz der unterschiedlichen Eingangseignung für Fachunterschiede halten, müssten die zulassungsbeschränkten Studiengänge bessere Noten aufweisen, als die Studiengänge mit gleichem Abschluss ohne Numerus Clausus. Psychologie und Biologie, die klassische NC-Fächer darstellen, weisen auch tatsächlich die besten Noten unter den Diplomstudiengängen auf. Allerdings sind die Noten in Physik, Mathematik und Chemie, allesamt Fächer ohne vergleichbar strenge und langjährige bundesweite NC-Historie ähnlich gut oder nur unbedeutend schlechter als die Noten in Psychologie und Biologie. Dass die Noten in Medizin mit Abschluss Staatsexamen, ebenfalls traditionell zulassungsbeschränkt, schlechter sind als in den meisten Diplom- und Magisterstudiengängen, belegt zudem, dass die Unterschiede im Notenniveau zwischen Abschlussarten nicht allein durch unterschiedliche Eingangseignungen erklärt werden können (vgl. Abb.12-14).

Sollten unterschiedliche Leistungsniveaus zwischen Fächern, Abschlüssen oder Hochschulen zu Beginn oder auch im Laufe des Studiums durch die Zusammensetzung der Studierenden entstehen, könnten diese Unterschiede möglicherweise auf soziodemographische Variablen zurückgeführt werden. Towfigh et al. (2014) weisen hierzulande etwa schlechtere Ergebnisse in juristischen Probeklausuren und in der Examensnote für Frauen und Prüflinge mit einem Namen, der auf einen Migrationshintergrund schließen lässt, nach. Van den Bussche et al. (2006) finden in der ärztlichen Vorprüfung einen Zusammenhang zwischen steigendem Anteil Ausländer*innen an Fakultäten und steigendem Notendurchschnitt– allerdings keinen geschlechtsspezifischen Effekt. Auch für wirtschaftswissenschaftliche Studiengänge finden Studien an deutschen Hochschulen keine signifikanten Unterschiede zwischen den Noten von Männern und Frauen (Giese et al. 2013; Ottwaska 1971). Der Einfluss der sozialen Herkunft bleibt für die hiesige Notengebung aufgrund der wenigen und nicht eindeutigen Befunde unklar: Ottwaska (1971) findet keinen signifikanten Einfluss des sozialen Hintergrunds, operationalisiert über den Beruf des Vaters, auf die Noten in den Wirtschaftswissenschaften an der Universität Mannheim. Bei Hampe (1977; 1978) finden sich an der Universität Marburg in Jura leicht bessere Examensnoten für Prüflinge, deren Vater einen akademischen Abschluss aufweist, während in VWL, Medizin und Zahnmedizin Prüflinge ohne entsprechende Herkunft leicht bessere Notendurchschnitte aufweisen.

Dass es eine fach- und/oder abschlussspezifische Lehrqualität gibt, die sich auf das Notenniveau einzelner Fächer und Abschlüsse auswirkt, wurde bisher noch nicht untersucht. Eine hochschulübergreifend gleiche Lehrqualität scheint aber eher unwahrscheinlich - wahrscheinlicher ist es, dass innerhalb von Studiengängen eine unterschiedlich hohe Lehrqualität an einzelnen Hochschulen vorzufinden ist. Hierfür spricht etwa der Befund von van den Bussche et al. (2006), die zeigen dass bei besserer Personalausstattung bessere Noten in der ärztlichen Vorprüfung zustande kommen.

Zur Erklärung von unterschiedlichen Bewertungsstandards, also *leistungsunabhängigen* Unterschieden zwischen Abschlüssen, Fächern und Hochschulen, existieren verschiedene Ansätze. Sie beziehen sich meist auf Unterschiede zwischen Fächern, in etwas geringerem Umfang auf Hochschulunterschiede in der Notengebung. Der Großteil der vermuteten Mechanismen lässt sich jedoch auch auf abschlussspezifische Notenunterschiede übertragen.

In der Regel werden in der Begründung von Notenunterschieden epistemologische (für Unterschiede im Notenniveau zwischen Fächern) oder (wissenschafts-)organisatorische Unterschiede zwischen Abschlüssen, Fächern oder Hochschulen herangezogen. So führen Goldman und Hewitt (1975) ebenso wie Strenta und Elliot (1987) nach Kontrolle von Leistung bestehende Notenunterschiede zwischen Fächern auf unterschiedliche Formen der Wissensakkumulation und -organisation zurück. Ihre Argumentation stützt sich auf die Annahme eines Selektionsprozesses, der entgegen der Logik der oben erwähnten Möglichkeit besserer Noten aufgrund der leistungsstärkeren Studierenden verläuft. Naturwissenschaftliche Fächer, in denen die Noten an US-amerikanischen Colleges und Universitäten am schlechtesten sind, so die Autor*innen, sind hierarchisch strukturiert und verlangen im Gegensatz etwa zu sozialwissenschaftlichen Fächern eine stufenweise Aneignung von Wissen. Studierende, die in den frühen Stufen der Wissensaneignung scheitern, können in weiterführenden Kursen demnach aufgrund des fehlenden Basiswissens keine guten Leistungen mehr erbringen. In den naturwissenschaftlichen Fächern, so die Argumentation der Autor*innen weiter, existiert ein größerer Anteil an Faktenwissen als etwa in den Sozialwissenschaften. Dadurch würden Verständnisprobleme bei Studierenden eher sichtbar werden. Im Endeffekt sei so in den Naturwissenschaften ein höherer Mindestlevel an Fähigkeiten nötig, um einen bestimmten Notendurchschnitt zu erreichen, was ein Ausweichen der weniger befähigten oder engagierten Studierenden in nicht-naturwissenschaftliche Fächer bedinge (Goldman/Hewitt 1975; Strenta/Elliot 1987).

Dass in den nicht-naturwissenschaftlichen Fächern mit dem geringeren Leistungsstandard dennoch die besseren Noten vergeben werden, erklären sie anhand der Adaptation-Level-Theory. Sie besagt, dass Wahrnehmungen immer an einem Bezugspunkt orientiert sind und Wahrgenommenes immer hinsichtlich des Abstands zu diesem Bezugspunkt eingeordnet wird. Der Bezugspunkt selbst ist dabei nicht fix, sondern wird durch die individuellen Erfahrungen determiniert (Helson 1947, 1948). Übertragen auf die Benotung bedeutet dies, dass studentische Leistung immer relativ zum durchschnittlichen Leistungslevel einer Prüfungsgruppe, die einen Hintergrundstimulus darstellt, wahrgenommen wird. Befinden sich viele leistungsstarke Studierende in einer Gruppe, ist der Bezugspunkt für die zu bewertende Leistung höher, als wenn viele leistungsschwache Studierende in einer Gruppe auftreten. Verschiebt sich der Bezugspunkt der zu bewertenden Leistung aufgrund eines veränderten Hintergrundstimulus (einer schwächeren Durchschnittsleistung) nach unten, ist es leichter, die Leistungserwartungen zu erfüllen und so eine gute Note zu erzielen. Auch wenn diese Argumentation

von den Autor*innen lediglich auf Fachunterschiede bezogen wird, ist sie prinzipiell auch auf Abschluss- und Hochschulunterschiede übertragbar, ließe sich begründen, durch welche Merkmale bestimmte Abschlussarten oder bestimmte Hochschulen eher leistungsstarke, andere eher leistungsschwache Studierende anziehen.

Diese Variante der Adaption Theory ist für die Notenunterschiede zwischen Fächern an deutschen Hochschulen allerdings nicht ohne weiteres übertragbar - hier werden, wie oben dargestellt, in den Naturwissenschaften im Durchschnitt sehr gute Noten vergeben. Auch, dass der beschriebene Mechanismus an deutschen Hochschulen ausbleibt und eine Konzentration der leistungsstärkeren Studierenden in den Naturwissenschaften und der leistungsschwächeren Studierenden in nichtnaturwissenschaftlichen Fächern in leistungskonformer Weise zu besseren Noten für Erstere und nicht für Letztere führt, scheint wie bereits zuvor dargestellt, nicht der Fall zu sein. Gegen die These, in den Naturwissenschaften würden sich vorwiegend leistungsstarke Studierende sammeln, während leistungsschwache Studierende in andere Fächer ausweichen, spricht zudem, dass diese Annahme impliziert, Studienerfolg beruhe allein auf einer allgemeinen Studierfähigkeit, die fachunabhängig in Leistung im Studium umgesetzt werden kann. Das würde bedeuten, dass Studierende, die etwa in Chemie, Mathematik oder Physik starke Leistungen erbringen, auch in Fächern wie Geschichte, Soziologie oder Germanistik zwangsläufig in der Lage wären, gute Leistungen abzuliefern, während dies im umgekehrten Fall nur begrenzt möglich wäre.

Empirisch zeigt der hohe Vorhersagegehalt der Gesamtabiturnote bei gleichzeitig schwächerer Prognosekraft der Einzelfachnoten zwar, dass Studienerfolg in der Tat zum Teil auf fachunabhängigen Kompetenzen beruht - Studierende müssen eine grundsätzliche Bereitschaft und grundsätzliche Fähigkeiten mitbringen, die es ihnen erlauben, sich selbst zu organisieren und ihr Studium erfolgreich als Projekt zu leiten (Huber 1994; Köller/Baumert 2002). Problematisch ist allerdings, dass selbst das größte Ausmaß an generellen Kompetenzen niemanden weiter bringt, wenn fachspezifische Kompetenzen nicht ausreichend internalisiert und angewendet werden können. Hier gilt es zu bedenken, dass die Vorhersagekraft der Abiturnote nur unter der Bedingung eines zuvor in Form der Studienfachwahl erfolgten Selbstselektionsprozesses Gültigkeit besitzt. Die Annahme einer generellen Studierfähigkeit, die fachunabhängig zu gleichem Studienerfolg führt wurde entsprechend bereits empirisch widerlegt (etwa Köller/Baumert 2002; zur historischen Entwicklung der Diskussion um die Existenz einer generellen Studierfähigkeit siehe Lewin/Lischka 2004) und ist nur noch eingeschränkt in Bezug auf grundsätzliche Erfolgsdeterminanten haltbar. Dies entspricht der in Kapitel 3 aufgezeigten großen Bandbreite an Lernanforderungen innerhalb des Fachspektrums in Kombination mit der Übereinstimmung zwischen fachlichen Anforderungen und individuellen Dispositionen (Armingeon 2001; Georg 2005; Windolf 1992).

Auch andere Forscher*innen können einen Zusammenhang zwischen Unterschieden im Notenniveau und Selektionsprozessen erkennen. Im Gegensatz zu den beiden bisher vorgestellten Annahmen, dass Selektionsprozesse die Ursache unterschiedlicher Notenniveaus sind, wird jedoch häufig auch die Möglichkeit in Betracht gezogen, dass Noten als Steuerungsinstrument eingesetzt werden, um Wanderungsprozesse zwischen Hochschulen oder Fächern zu kontrollieren. Achen und Courant (2009) wie auch Bagues et al. (2008) und De Paola (2008) vermuten beispielsweise, dass Fakultäten bzw. einzelne Lehrende (Dickson 1984) mit niedrigen Studierendenzahlen dazu tendieren, bessere Noten zu vergeben, um Studierende in ihre Kurse zu locken, unter anderem, um die mit höheren Studierendenzahlen einhergehenden höheren Fördermittel zu erhalten oder einfach, um ihren Arbeitsplatz zu sichern. Diese Vermutung wird durch den Befund erhärtet, dass Studierende mit guten Abschlussnoten zum Teil geringeren Erfolg auf dem Arbeitsmarkt aufweisen, als Absolvent*innen mit vergleichsweise schlechteren Noten (Bagues et al. 2008).

Ursächlich für die Unterschiede im Notenniveau wären demnach unterschiedlich starkes studentisches Interesse an einzelnen Fächern oder Abschlüssen bzw. unterschiedlich attraktive Hochschulstandorte, was aus Sicht einzelner Hochschulakteur*innen ein Gegensteuern erfordert. Ähnlich argumentiert Freeman (1999), der die Argumentation allerdings noch um einen Schritt erweitert. Er sieht fachspezifische Entlohnungsstrukturen des Arbeitsmarkts als eigentliche Ursache fachspezifischer Notenniveaus an. Ausgangspunkt seiner Argumentation ist sein Befund, dass Studierende aus Fächern mit guten Noten im Erwerbsleben durchschnittlich geringere Verdienste erzielen als Studierende aus Fächern mit schlechteren Noten. Unter der Annahme, dass die späteren Erwerbschancen einen nennenswerten Grund für die Entscheidung, ein Studium zu absolvieren darstellen, würde sich nach Freeman ohne ausgleichende Steuerung ein Großteil der Studierenden in Fächer einschreiben, von denen bekannt ist, dass sie die Wahrscheinlichkeit auf ein hohes Einkommen im späteren Beruf erhöhen. Über die Notengebung wird eine Ausgleichssteuerung möglich: Studierende, die sich für Fächer mit vergleichsweise schlechten Einkommensaussichten entscheiden, werden dafür mit guten Noten belohnt.

Freemans Argument ist allerdings so stark auf einen einzelnen ökonomischen Faktor beschränkt, dass in dieser Logik die durch ein Studium möglichen Verdienstmöglichkeiten als einziger oder zumindest als hauptsächlicher Grund für die Aufnahme eines Studiums unterstellt werden. Die Tatsache, dass viele Studierende - und zwar gerade die in Fächern ohne sehr hohe Verdienstmöglichkeiten häufig eine intrinsische Motivation zum Studium aufweisen (Ramm et al. 2011; Wilke 1976), lässt starke Zweifel an dieser impliziten Grundlage von Freemans These aufkommen.

Dass der Arbeitsmarkt Einfluss auf die Notengebung hat, vermuten auch Hitpass und Trosien (1987). Sie ziehen die Möglichkeit in Betracht, dass sich Fachunterschiede aus unterschiedlichen Arbeitsmarktchancen für bestimmte Fächer/Abschlüsse ergeben. Der Mechanismus, den die Autoren ver-

muten, beinhaltet die Wahrnehmung abschluss- bzw. fachspezifischer Arbeitsmarktchancen durch die Prüfenden, die sie dazu bewegen könnte, Nachteile auf dem Arbeitsmarkt durch gute Noten auszugleichen. Allerdings können Müller-Benedict und Tsarouha (2011) zeigen, dass ein erhöhtes Angebot an Absolvent*innen auf dem Arbeitsmarkt, also erhöhte Prüfungszahlen auch erhöhte Durchfallquoten mit sich bringen, was gegen die These einer milderen Benotung bei schlechterer Arbeitsmarktlage spricht. Möglich ist jedoch nach den Autor*innen, dass genau der gegenteilige Effekt eines überfüllten Arbeitsmarktes eintritt und Prüfende bei Wahrnehmung eines Überflusses an Absolvent*innen die Selektionsfunktion der Hochschule ernster nehmen als gewöhnlich. Dieses von Müller-Benedict und Tsarouha vorgebrachte Argument, um mögliche Zyklen in der Notengebung zu erklären, lässt sich auch fach- oder (im Falle von Lehramtsnoten) abschlussspezifisch anwenden, sind entsprechende Unterschiede in den Arbeitsmarktchancen festzustellen.
Solange die Unterschiede in den Arbeitsmarktchancen zwischen den Fächern konstant bleiben, könnten dadurch auch konstante Unterschiede im Notenniveau erklärt werden. Änderungen im Verhältnis zwischen den Arbeitsmarktchancen der Fächer müssten dann auch mit Verzögerungen zu Veränderungen im Verhältnis der Notenniveaus führen. Im Falle regionaler Arbeitsmarktunterschiede ist auch ein unterschiedlich starker Einfluss des Arbeitsmarktes auf die Noten einzelner Hochschulen im gleichen Fach denkbar, sollten die Prüfenden eher diese Größe als die nationale Entwicklung wahrnehmen.

Die Selektionsneigung von Prüfenden kann, wie bereits erläutert wurde, durch verschiedenste Faktoren beeinflusst werden. So könnten auch die lokalen Lehr- und Lernbedingungen unabhängig von der Leistung auf die Bewertung wirken, indem sie die Selektionsneigung der Prüfenden beeinflussen. Hier sehen Müller-Benedict und Tsarouha unter Verweis auf Hitpass/Trosien (1987) die Möglichkeit, dass schlechte Rahmenbedingungen, zum Beispiel überfüllte Veranstaltungen, durch gute Noten wiedergutgemacht werden könnten. Allerdings stellt Krempkow (2006; auch Baird 2009; Koedel 2011 und Kokkelenberg et al. 2008) fest, dass Studiengänge mit besseren Abschlussnoten auch geringere Studierendenzahlen aufweisen, was eher in Einklang mit den oben angeführten Argumentationen steht, die gute Noten als Lockmittel für mehr Studierende verstehen. Auch die Annahme leistungskonformer Verbesserung der Noten aufgrund besserer Leistung durch günstigere Betreuungsrelationen in kleineren Kursen passt hierzu (Chapman 1994).
Neben den Lehr- und Lernbedingungen spielen auch die konkreten Prüfungsbedingungen eine Rolle – sie können sich sowohl informal als auch formal je nach lokalem Kontext deutlich unterscheiden (etwa Knight 2006), was möglicherweise unterschiedliche Notenniveaus begünstigt. Hier wird in der Literatur etwa auf unterschiedliche Wahlmöglichkeiten in verschiedenen Studiengängen (Barth et al. 2009) oder auf unterschiedlich breite Kursangebote (in Bezug auf Notenverbesserungen: Falkenberg

1996) hingewiesen: Hier sind die Annahmen, dass Studierende bessere Ergebnisse erzielen, je mehr Kurse sie gemäß ihrer Interessen und nicht aus Pflichtvorgaben wählen und dass leichter gute Ergebnisse erzielt werden, wenn sich die Kenntnisse und Fähigkeiten von bestimmten Kursen (oder Teilprüfungen) in andere übertragen lassen.

Kolevzon (1981) berichtet zudem, dass auch die Art der eingesetzten Prüfungsverfahren und die Einstellungen der Lehrenden zur jeweiligen Prüfungspraxis zwischen Fächern mit hohen Verbesserungen im Notenniveau und solchen mit geringen Verbesserungen variieren - in ersteren werden häufiger mündliche Prüfungen und Aufsätze gefordert, in letzteren häufiger standardisierte Tests durchgeführt.

Hitpass und Trosien (1987), wie auch Müller-Benedict und Tsarouha (2011), stellen die These auf, dass die Zusammensetzung der Prüfungskommission einen Einfluss auf Unterschiede im Notenniveau zwischen Abschlüssen hat. Hier wird die Rolle von externen Prüfenden als Garant*innen für eine hohe Qualität betont, die sich in schlechteren Noten für Staatsexamensprüfungen als für Diplom- und Magisterprüfungen niederschlägt. Generell heben Müller-Benedict und Tsarouha die Rolle der Akteur*innen in der Notengebung, den Hochschullehrenden, hervor. Auch wenn die Lehrenden im Falle struktureller systematischer Unterschiede nur die Funktion eines Mediums zwischen strukturellem (leistungsexternen) Input und zugehörigem Notenoutput darzustellen scheinen, sind sie doch die zentrale Schnittstelle, an der leistungsabhängige und -unabhängige Faktoren in ein Prüfungsergebnis in Form einer Note umgewandelt werden.

Entsprechend setzen einige Ansätze zur Erklärung von unterschiedlichen Notenniveaus bei der Rolle der Prüfenden an. Achen und Courant (2009) weisen etwa darauf hin, dass Prüfende in unterschiedlichen Fächern und auch innerhalb dieser in unterschiedlichen Kursen unterschiedlich stark formalisierte Prüfungs- und Bewertungsverfahren einsetzen. Sie zeigen, dass die Noten in den Fächern (und Kursen) schlechter sind, in denen diese Verfahren intransparenter und dadurch leichter anzufechten sind. Sie folgern daraus, dass Prüfende in Fächern mit hoch standardisierten Prüfungs-und Bewertungsverfahren geringere Zeit und Mühe dafür aufwenden müssen, sich für schlechte Noten gegenüber Studierenden zu verteidigen und sie sich deshalb eher „erlauben“ können, als solche, die hierfür einen größeren Aufwand investieren müssten und deshalb Beschwerden durch gute Noten vorbeugen. Auch unterschiedliche Bezugsnormenorientierungen der Lehrenden verschiedener Fächer könnten, wie in Kapitel 2 erläutert, unterschiedliche Notenniveaus theoretisch begünstigen, sollte in einigen Fächern (vorwiegend) die sachliche, in anderen die soziale Bezugsnorm verwendet werden.

Und auch die Frage, wie sich die durch gute Benotungen gesparte Zeit und Mühe verwenden lässt, gibt möglicherweise Aufschluss über die Entstehung und Stabilität fach- und hochschulspezifischer Notenniveaus. So lässt sich die von Franz (2010) vorgetragene Argumentation zur Erklärung der generellen Verbesserung von Noten im Zeitverlauf, Prüfende geben gute Noten, um die durch Vermei-

dung von Beschwerden gesparte Zeit auf Forschung zu verwenden, auch auf unterschiedliche Bewertungsstandards zwischen Fächern und Hochschulen übertragen. Die These wäre dann, dass Prüfende aus forschungsintensiveren Fächern oder an forschungsaktiveren Hochschulen bessere Noten vergeben als ihre Kolleg*innen, die weniger Forschung betreiben. Auch fachspezifische Unterschiede in der Höhe der Prüfungsbelastung könnten im Hinblick auf einen möglichen Zusammenhang zwischen der Notengebung und dem Zeitbudget eine Rolle spielen, wenn bei hoher Prüfungsbelastung ein milderes Selektionsklima dazu genutzt wird, den Aufwand für die Prüfungen durch die Verringerung von Beschwerden auszugleichen.

Ferner wird die Zusammensetzung des Lehrkörpers und der Studierenden in Hinblick auf Arbeitsverhältnis, auf normative Einstellungen oder auch auf den sozialen Hintergrund als mögliche Ursache für bestimmte Notenniveaus in Betracht gezogen. Barth et al. (2009) sowie Moore und Trahan (1998) präsentieren den Befund, dass befristete Lehrkräfte (non-tenure track) bessere Noten geben als festangestellte Professor*innen (tenured/tenure-track) und vermuten als Ursachen hierfür den erhöhten Druck, gute Arbeit abzuliefern, um einen neuen Vertrag zu erhalten sowie mangelnde Erfahrung in der Notengebung. Gohmann und McCrickard (2001), Kezim et al. (2005) und Sonner (2000) können diesen Befund für Privatdozierende bestätigen, wobei Gohmann und McCrickard wie auch Kezim et al. im Gegensatz zu Moore und Trahan keine signifikanten Unterschiede in der Höhe der vergebenen Noten zwischen festangestellten und befristet angestellten Professor*innen finden können. Doss et al. (2005a-f) weisen nach, dass in Teilzeit angestellte Lehrende in fünf von sechs Kursen an einer von ihnen untersuchten Hochschule bessere Noten vergeben als Vollzeitbeschäftigte. Bar und Zussmann (2012) zeigen, dass demokratisch wählende Prüfer*innen die Notenskala hinsichtlich über SAT-Scores festgestellter Leistungsunterschiede nicht in gleichem Maße ausreizen wie republikanisch wählende Prüfer*innen, in deren Leistungsbeurteilungen sich die Bandbreite der Ergebnisse der Eignungstests eher widerspiegelt. Prüfer*innen, die republikanisch wählen, bewerten zudem ethnische Minderheiten schlechter als demokratische Prüfer*innen. Jewell und McPherson (2012) schließlich präsentieren Daten, nach denen weibliche Lehrende in den USA bessere Noten geben als männliche, während die Ethnizität der Lehrenden keinen Effekt aufweist.

Sollten systematische Unterschiede in den Arbeitsverhältnissen oder den politischen Einstellungen von Dozierenden zwischen verschiedenen Fächern oder auch regional zwischen Hochschulstandorten bestehen, ist es demnach möglich, dass diese Besonderheiten Einfluss auf die Höhe des Notenniveaus ausüben. Auch die Interaktion mit Studierenden und die Einstellungen ihnen gegenüber variieren möglicherweise zwischen den Fächern, was bezüglich eines Zusammenhangs von Sympathie und Benotung – bis hin zu Diskriminierung - relevant sein könnte. Die Einstellungen der Lehrenden würden sich dann in Abhängigkeit der sich ihnen entgegenstehenden Zusammensetzung der Studierenden auswirken. An dieser Stelle sei etwa auf den eingangs erwähnten Befund schlechterer Noten für

weibliche Prüflinge und solche mit Migrationshintergrund ausgerechnet in den Rechtswissenschaften verwiesen – wobei ein derartiger, auf Diskriminierung abstellender Zusammenhang ohne einen empirischen Beleg natürlich rein spekulativer Natur bleibt.

Die Zusammenfassung der Erklärungsansätze zu Unterschieden im Notenniveau zwischen Abschlussarten, Fächern und Hochschulen zu gegebenen Zeitpunkten oder über einen längeren Zeitraum gemittelt verdeutlicht, dass in der Literatur eine Vielzahl von möglichen Einflussfaktoren und Mechanismen diskutiert wird. Sie lassen sich bei näherer Betrachtung teilweise untereinander ergänzen oder verbinden, teilweise werden ihnen aber auch gegensätzliche Wirkungen zugesprochen. So können etwa die Arbeitsmarktchancen oder auch die Lehrqualität theoretisch sowohl eine Verbesserung als auch eine Verschlechterung der Noten in einzelnen Fächern, in einzelnen Abschlussarten oder an einzelnen Hochschulen bewirken. Lediglich epistemologische Unterschiede zwischen den Fächern können bedingungslos langfristig stabile Unterschiede im Notenniveau produzieren, der Einfluss aller anderen theoretisch wirksamen Faktoren auf das Notenniveau einzelner Untersuchungseinheiten hängt im Zeitverlauf immer von der eigenen Entwicklung ab. So können bessere Noten in Fach A als in Fach B beispielsweise überhaupt nur für den Zeitraum auf unterschiedliche Rahmenbedingungen in der Lehre zurückgeführt werden, in dem die Lehrbedingungen in Fach A nachweislich besser sind als in Fach B (Hypothese der besseren Lehrqualität) oder umgekehrt (Ausgleichshypothese).
Bisher wurden nur einzelne dieser potentiellen Einflussfaktoren auch empirisch überprüft. Und wenn, dann wurden sie, ebenso wie die jeweils gewählte Vergleichsebene (Abschlüsse, Fächer, Hochschulen), in der Regel isoliert betrachtet, ohne mögliche Wechselwirkungen oder Abhängigkeiten zwischen ihnen zu ergründen.

Es zeigt sich dabei, dass ein Zusammenhang zwischen der Eingangseignung der Studierenden und ihren späteren Prüfungsleistungen besteht, unterschiedliche Eingangseignungen aber weder auf Fach-, noch auf Abschluss- oder Hochschulebene zur Erklärung der nachgewiesenen Unterschiede im Notenniveau ausreichen. Für Deutschland sind Hinweise auf soziodemographisch bedingte Notenunterschiede vorhanden, die jedoch nicht für eine umfassende Einordnung möglicher Effekte ausreichen. Der Einfluss der Lehrqualität ist bisher nicht umfassend untersucht worden, für die ärztliche Vorprüfung ist jedoch ein Zusammenhang zwischen Noten und Personalausstattung belegt.
Hinsichtlich leistungsexterner Einflüsse deutet sich an, dass Prüfende arbeitsmarktabhängige Selektionsneigungen aufweisen und dabei eher mit härterer Selektion auf schlechte Arbeitsmarktlagen reagieren als mit besonderer Milde. Dass die Prüfungsbelastung, die formalen Prüfungsbedingungen, die eingesetzten Prüfungsverfahren, der Grad der Standardisierung und die Zusammensetzung der Lehrenden die Notengebung systematisch beeinflussen, legen Befunde aus den USA nahe. Der Stand der

Forschung stützt damit Forschungshypothese FH1 (leistungskonforme *und* leistungsexterne Faktoren verursachen Unterschiede im Notenniveau). Dass es mildere Noten als Ausgleich für schlechtere Lehrbedingungen gibt, ist hingegen nicht nur für Deutschland sondern auch international ebenso wenig empirisch abgesichert wie Einflüsse der Zusammensetzung der Prüfungskommissionen oder der Forschungsintensität, die durchaus denkbar sind.

Unwahrscheinlich ist, dass in Deutschland eine Hierarchie in grundsätzlichen, epistemologisch begründbaren Anspruchsniveaus zwischen den Fächern besteht, die Unterschiede im Notenniveau produziert, da die dieser These zugrunde liegende Annahme einer generellen Studierfähigkeit nur begrenzt plausibel ist. Unabhängig von einer Hierarchisierung der Anforderungen entlang epistemologischer Merkmale bleibt jedoch die Möglichkeit bestehen, dass epistemologische Merkmale den Grad der Standardisierung in Prüfungen beeinflussen. Dass Noten genutzt werden, um Studierende anzuziehen und dadurch finanzielle Mittel aufzustocken ist für deutsche Hochschulen aufgrund der selbst in heutigen Zeiten des New Public Management immer noch geringen Kopplung von Fördermitteln an derartige Outputfaktoren ebenfalls nicht anzunehmen (Bauer/Grave 2011). Tabelle 2 ordnet die in der Literatur besprochenen Ursachen für unterschiedliche Bewertungsstandards, also für leistungsexterne Gründe für Differenzen im Notenniveau, den zuvor hergeleiteten Ursachenkategorien zu.

Tabelle 2: Einordnung der besprochenen Ursachen für unterschiedliche Bewertungsstandards in Ursachenkategorien

In der Literatur angeführte Ursachen für unterschiedliche Bewertungsstandards	Übergeordnete Ursachenkategorien
- Wahlmöglichkeiten, - Breite des Kursangebots - Zusammensetzung der Prüfungskommission	- unterschiedliche formale Prüfungsbedingungen
- Formalisierung	- unterschiedliche Standardisierungsgrade
- Gute Noten, um mehr Zeit zum Forschen zu haben	- unterschiedliche Forschungsintensität
- Gute Noten, um hohe Prüfungsbelastung zu verringern	- unterschiedlich hohe Prüfungsbelastung
- Gute Noten als Ausgleich schlechter Lehrbedingungen	- unterschiedliche Rahmenbedingungen für Lehre
- Arbeitsverhältnis, Politische Einstellungen, Geschlecht	- unterschiedliche Zusammensetzung der Lehrenden
- Noten als Steuerungsinstrument aufgrund ungleicher Einkommensaussichten - Noten zur verschärften Selektion bei Überfüllung auf dem Arbeitsmarkt	- unterschiedlich gute Arbeitsmarktchancen der Prüflinge
- Noten als Steuerungsinstrument, um Förderung zu erhöhen - Noten als Steuerungsinstrument zur Arbeitsplatzsicherung	- unterschiedliche Finanzierungsstrukturen

6.3 Die Entwicklung von Noten im Zeitverlauf

Der Vergleich von Notenniveaus zu gegebenen Zeitpunkten offenbart erklärungsbedürftige Unterschiede in der Notenvergabe zwischen Untersuchungseinheiten verschiedener Vergleichsebenen. Dass diese Unterschiede auch gemittelt über bestimmte Zeitabschnitte nachzuweisen sind, offenbart eine kurz- bis mittelfristige Stabilität. Zuverlässige Aussagen über die konkrete Entwicklung eines

Notenniveaus können jedoch auch anhand von Daten, die es zu mehreren Zeitpunkten abbilden, nicht getroffen werden, solange keine kontinuierlichen Messreihen vorliegen. Gemeinsamkeiten oder Unterschiede in der Entwicklungsdynamik der Noten können anhand von Querschnittsanalysen nicht erkannt werden. Dies ist nur durch die Einnahme einer Längsschnittperspektive möglich, durch die sowohl ein intra-individueller Vergleich einzelner als auch ein inter-individueller Vergleich der Entwicklung mehrerer Untersuchungseinheiten erfolgen kann.

6.3.1 Grade Inflation – Notenverbesserung als Problem

Die empirische Forschung zur langfristigen Entwicklung von Hochschulnoten widmet sich vor allem der Analyse eines Phänomens, der sogenannten ‚grade inflation'; ein Thema, das vor allem in den USA eine lange Tradition besitzt. Der Begriff grade inflation beschreibt eine langfristige, kontinuierliche Verbesserung des Notenniveaus ohne eine entsprechende Verbesserung der Prüfungsleistung (Bejar/Blew 1981). Als Analogie zum Entwertungsprozess von Geld als Währung im Falle einer Preisinflation, wird Noten im Falle der grade inflation ein Wertverlust attestiert, in dem Sinne, dass eine bestimmte Note nicht mehr die gleiche Leistung widerspiegelt, die zu einem früheren Zeitpunkt benötigt wurde, um diese Note zu erhalten (Birnbaum 1977).

Einer Noteninflation werden mehrere negative Auswirkungen zugeschrieben. So wird die Signalkraft von Noten als Abbildung erbrachter Leistung durch eine Zunahme guter Noten ohne gleichzeitige Zunahme von Leistung als gefährdet betrachtet (Chan/Hao/Suen 2007; Pressman 2007). Daraus wird häufig eine eingeschränkte Vergleichbarkeit der Leistung von Absolvent*innen sowohl im Zeitvergleich als auch untereinander im Querschnittsvergleich gefolgert (etwa Mitchell 1998; Wilson 1999). Als negative Konsequenzen einer verringerten Signalkraft von Noten werden dann zum Beispiel eine steigende Bedeutung askriptiver Merkmale wie soziale Herkunft oder Ethnizität als Kriterien zur Leistungsbeurteilung, etwa durch Arbeitgeber*innen, befürchtet (Schwager 2008). Gute Noten für geringe Leistung senken zudem das Engagement von Studierenden (Babcock 2010) und sorgen für verzerrte Entlohnungsstrukturen, da vor allem leistungsschwächere Studierende von sinkenden Bewertungsstandards für gute Noten profitieren (Mc Spirit/Jones 1999). Auch werden Studierende aus Hochschulen mit hoher Noteninflation an Übergängen im Bildungswesen (Master, Promotion) strukturell bevorteilt (Ogilvie/Jelavic 2013).

Koedel (2011) nimmt zudem an, dass zu viele gute Noten im Studium verhindern, dass Studierende es frühzeitig erkennen, wenn sie für bestimmte Berufe nicht geeignet sind, Stone (1995) sieht mit einer Entwertung von Noten finanzielles Missmanagement einhergehend und stuft Noteninflation als Betrug an der Gesellschaft ein.

Gelegentlich wird in der Literatur eine Schwerpunktsetzung auf eine spezifische Folge der Noteninflation gesetzt, die ‚grade compression' (etwa Ostrovsky/Schwarz 2003). Ausgangspunkt für diese Schwerpunktsetzung ist die Annahme, dass Noteninflation an sich ein vergleichsweise geringes Prob-

lem darstellt, solange die Noten aller Studierenden sich im stabilen Verhältnis zueinander verbessern. Eine Differenzierung zwischen den Leistungen der Prüflinge im Querschnitt ist dann nicht gefährdet, solange nur der Notendurchschnitt sinkt. Relevant ist lediglich dass die Relation derjenigen, die zu einem früheren Zeitpunkt ein „befriedigend" erhielten zu denen, die ein „gut" erhielten der Relation derjenigen, die zu einem späteren Zeitpunkt ein „gut" erhalten zu denen, die ein „sehr gut" erhalten, entspricht. Erst, wenn ein gewisser Anteil Absolvent*innen mit der Bestnote abschließt und sich so nicht weiter verbessern kann, während ein immer größer werdender Anteil der Prüflinge mit schlechterer Leistung trotz eines konstanten Leistungsgefälles im Zeitverlauf durch eine inflationäre Notenvergabe ebenfalls näher an die Bestnoten heranrückt, wird die differenzierende Aussagekraft von Noten *zwangsläufig* eingeschränkt (Hu 2005).

6.3.2 Empirische Befunde zur Notenverbesserung in Nordamerika

Als Startpunkt für den postulierten Entwertungsprozess von Noten werden in den USA die 1960er Jahre betrachtet. Juolas Veröffentlichung zur Entwicklung von College Noten zwischen 1960 und 1973 gilt als „first scholarly paper to make a statistically credible case for a national epidemic of grade inflation" (Kamber 2008:52), ein Verweis auf diese Untersuchung fehlt in kaum einer Übersicht, die sich der Entwicklung von grade inflation in den USA widmet. Juola stellt eine Verbesserung der Durchschnittsnoten von Undergraduate-Studierenden an 134 Colleges um 0.404 Noten im untersuchten Zeitraum fest (Juola 1976). Ein Vergleich mit Noten der Michigan State University ab 1941, für die er keine entsprechenden Verbesserungen des Notenniveaus feststellen kann, lässt Juola schlussfolgern, dass es sich hierbei um ein neues Phänomen handelt (ebd.).

Seit Juolas Studie wurde die Entwicklung der Notenniveaus an Colleges und Universitäten in den USA immer wieder untersucht[41]. Rogers (1983) stellt in einer Analyse der jährlichen Abschlussnoten an der Universität Northern Iowas von 1929 bis 1981 fest, dass sich auch dort die Noten erst ab Beginn der 1960er Jahre stetig verbessern, während sie zwischen 1929 und 1960 deutlichen Schwankungen unterliegen. Die Notenentwicklungen an zwei Vergleichsuniversitäten, für die die Daten erst ab 1947 bzw. 1958 vorliegen, weisen eine ähnliche Entwicklung auf. Kolevzon (1981) berichtet von einer Verbesserung der Notendurchschnitte um 0.298 Noten zwischen den Jahren 1969/70 und 1975/76 an einer von ihm untersuchten Universität, Birnbaum (1977) stellt ebenfalls verbesserte Durchschnitte bei Studierenden der University of Wisconsin in 1974/75 gegenüber 1968/69 fest, so wie Prather et al. (1979) dies im Zeitverlauf von 1970 bis 1975 an einer öffentlichen Universität tun. Sabot und Wakeman-Linn (1991) ermitteln bessere Noten an acht von neun untersuchten Colleges und Universitä-

[41] Suslow (1976) stellt eine Verbesserung der Abschlussnoten von Undergraduate-Studierenden um 0.47 Noten im Zeitraum von 1960 bis 1974 fest und bestätigt damit (auch im Ausmaß) Juolas Ergebnisse. Seine Studie findet in der Literatur wohl vor allem deshalb weniger Beachtung, weil die Zahl der betrachteten Einrichtungen wesentlich niedriger liegt als bei Juola und zudem zwischen den einzelnen Zeitpunkten stark variiert (zwischen zwei Standorten 1960 und 15 1974).

ten in 1985/86 gegenüber 1962/63 und stellen fest, dass die Verbesserungen mit abnehmenden Standardabweichungen einhergehen, ein Indiz für grade compression. Mullen (1995) findet eine Verbesserung der Notendurchschnitte von Studierenden im ersten Jahr an der University of Missouri um 0.09 Noten zwischen 1987 und 1992.

Mc Spirit und Jones (1999) stellen in Regressionsanalysen einen signifikanten Einfluss des Jahres des Studienbeginns auf die Notenhöhe der Abschlussnoten an einer öffentlichen Universität im Zeitraum von 1983-1996, Kezim et al (2005) an der Business School eines privaten Colleges von 1983-2003 fest. Schon Kwon et al. (1997) berichten von einer Verbesserung der Abschlussnoten an einer Business School 1993/94 gegenüber 1983/84. Cluskey et al. (1997) finden Verbesserungen in den Kursnoten einer privaten Universität zwischen 1980 und 1995, Cheong (2000) am Arts and Sciences College der Universität von Hawaii. Kuh und Hu vergleichen selbst berichtete Durchschnittsnoten von Studierenden der Zeiträume 1984-1987 und 1995-1997 an insgesamt 184 Colleges und Universitäten (die Noten von 14 davon sind in beiden Zeiträumen berücksichtigt) miteinander und stellen ebenfalls eine Verbesserung der Noten über alle Studierenden hinweg gemittelt fest (Kuh/Huh 1999). Grove und Wasserman (2004) finden eine Verbesserung der durchschnittlichen Semesternoten um 0.11 Noten in 2002 gegenüber 1998 an einer privaten Universität, Mathies und Webber (2009) zwischen 1985 und 2004 an einer öffentlichen Universität um 0.69 Noten, einhergehend mit sinkenden Standardabweichungen. Potter et al. (2001) finden an einem Liberal Arts College eine Verbesserung der Semesternoten um eine halbe Note zwischen 1969 und 1999, Compton und Metheny (2000) zwischen 1988 und 1995 um 0.08 Noten in den Kursnoten[42].

Wongsurawat (2009) stellt in einer Analyse der Bewerbungen an US-amerikanischen Law Schools Verbesserungen in den durchschnittlichen Abschlussnoten der Bewerber*innen von 1995 über 2000 bis 2007 fest. Baird (2009) diagnostiziert sich verbessernde Notendurchschnitte der Studierenden an der University of California für den Zeitraum 1980-2008, Rush et al. (2009) am College für Veterinärmedizin der Universität Kansas zwischen 1985 und 2006. Mulvenon und Ferritor (2005) weisen eine kontinuierliche Verbesserung von Kursnoten an der University of Arkansas von 1992/93 bis 2003/04, Bar et al. (2009) am Cornell College von 1990 bis 2004, Achen und Courant (2009) an der University of Michigan von 1992 bis 2008 und Jewell und McPherson (2012) sowie Jewell et al. (2013) an der University of North Texas von 1984 bis 2005 nach. Lowe et al. (2008) analysieren ein bundesweites sample von College Studierenden und ermitteln eine Verbesserung von 0.23 Noten in der durchschnittlichen Abschlussnote von Bachelor-Absolvent*innen 2000/01 gegenüber 1993, Levine und Cureton (1998) verweisen anhand von ähnlichen Umfragen auf eine Zunahme von Bestnoten („A- or higher“) und Abnahme von durchschnittlichen Noten („C or below“) im Zeitraum von 1969 bis 1993.

[42] Bearden und Wolfe (2001) berichten in Reaktion auf die Untersuchung von Compton und Metheny von ähnlichen Ergebnissen, legen aber leider keine Zahlen vor.

Popov und Bernhardt (2010) fassen Abschlussnotendurchschnitte aus dem Datenpool der Homepage www.gradeinflation.com, die eine Vielzahl an Universitäten und Colleges umfasst, für die Jahre 1960, 1980 und 2000 an 20 ausgewählten Universitäten zusammen und zeigen, dass sich die Notendurchschnitte an allen Universitäten im Vergleich zu 1960 verbessert haben. Die auf der Homepage von Rojstaczer zusammengetragenen Daten zeigen 2001 über 29 Hochschulen gemittelt eine Verbesserung von 0.6 Noten zwischen 1967–2001 (Yorke 2008). Der letzte Stand der Homepage zeigt eine Verbesserung von ca. 0.3 Noten für den Zeitraum 1983 bis 2013 gemittelt über 88 Colleges und Universitäten an (www.gradeinflation.com am 29.03.2016). In einer weiteren Sammlung von Semesternoten aus diversen Quellen stellen Rojstaczer und Healy (2012) dar, dass der Anteil der Bestnote ‚A' an allen Noten zwischen 1940 und 2009 von 15% auf 43% gestiegen ist.

Tabelle 3: Übersicht über zentrale Studien, die langfristige Notenverbesserungen im US-amerikanischen Hochschulsystem feststellen

Studie	Zeitraum	Untersuchungseinheit(en)	Untersuchte Noten	Zentraler deskriptiver Befund
Juola (1976)	1960-1973	134 Colleges, Auswertung für verschiedene Kategorien	Abschlussnoten (Undergraduate students)	Verbesserung um 0.404 Noten (über alle Colleges gemittelt), am stärkstem ab 1968
Suslow (1976)	1960-1974	15 Universitäten/Colleges	Abschlussnoten (Undergraduate students)	Verbesserung um 0.47 Noten (über alle Einrichtungen gemittelt),
Birnbaum (1977)*	1968/69 vs. 1974/75	Universität von Wisconsin, Auswertung nach Fächern	Notendurchschnitte (Senior students)	Verbesserung um 0.49 Noten (über alle Fächer gemittelt), fachspezifische Verbesserung
Prather et al. (1979)	1970-1975	1 Öffentliche Universität, Auswertung nach Fächern	Kursnoten (Undergraduate students)	Anstieg von „A“s um ca.12%, Abnahme von „C“s um ca. 8% (über alle Fächer gemittelt), fachspezifische Verbesserung, Fächer ohne Verbesserung
Kolevzon (1981)	1969/70 vs. 1975/76	1 Universität, Auswertung nach Fächern	Notendurchschnitte (Undergraduate students)	Verbesserung um 0.298 Noten (über alle Fächer gemittelt), fachspezifische Verbesserung, auch Fächer ohne Verbesserung
Rogers (1983)*	1929-1981	Universität von Northern Iowa und 2 Vergleichsuniversitäten	Notendurchschnitte (Undergraduate students)	Northern Iowa: Verbesserung ab 1960 um knapp eine halbe Note, auch an den Vergleichsuniversitäten Verbesserung ab Beginn der 1960er
Sabot/Wakeman- Linn (1991)	1962/63 vs. 1985/86	9 Colleges und Universitäten	Kursnoten (Einführungskurse)	Verbesserungen in 8 von 9 Hochschulen, fachspezifische Verbesserung
Mullen (1995)*	1987-1992	Universität von Missouri	Notendurchschnitte (Freshmen)	Verbesserung um 0.09 Noten (über alle Fächer gemittelt)
Cluskey et al. (1997)	1980-1995	1 Private Universität, getrennte Auswertung für Accounting Kurse und Business College	Kursnoten	Verbesserung auf Universitätsebene, im Business College und in den Accounting Kursen
Kwon et al. (1997)*	1983/84 vs. 1993/94	McKendree College (nur Business School)	Abschlussnoten (Undergraduate Students)	Verbesserung um 0.141 Noten
Levine/Cureton (1998)	1969 vs. 1976 vs. 1993	Bundesweites sample (4900 Studierende)	Berichtete Notendurchschnitte (Undergraduate students)	Anstieg von „A- or higher“ von 7% auf 26%, Rückgang von „C or below“ von 25% auf 9% (über alle Fächer gemittelt)
Kuh/Hu (1999)	1984-87 vs. 1995-97	184 Colleges und Universitäten, Auswertung nach Fächergruppen und Hochschultypen	Berichtete, bis zur Befragung erhaltene Noten	Verbesserung um 0.27 Noten (über alle Hochschulen gemittelt), institutionen- und fächergruppenspezifische Verbesserung
Mc Spirit/Jones (1999)	1983-1996	1 Öffentliche Universität, Auswertung nach ACT-Score Gruppen	Abschlussnoten (Undergraduate students)	Verbesserung um durchschnittlich 0.021 Noten pro Jahr, ACT-Score-spezifische Verbesserung
Cheong (2000)*	1987-1999	Arts & Sciences College der Universität von Hawaii	Kursnoten (Undergraduate Kurse)	Verbesserung um 0.15 Noten (über alle Fächer gemittelt), fachspezifische Verbesserung, auch Fächer ohne Verbesserung
Compton/ Metheny (2000)	1988-1997	1 Liberal Arts College, Auswertung nach Kurslevel, Fakultäten und Instituten	Kursnoten (Undergraduate Kurse)	Verbesserung um 0.08 Noten (über alle Kurse gemittelt), Kurs-, Fakultäts-, Institutsspezifische Verbesserung, auch Fakult./Institute ohne Verbesserung
Potter et al. (2001)*	1969-1999	1 Liberal Arts College (1 Mathematik-Kurs separat)	Semesternoten (Erstsemester)	Verbesserung um 0.5 Noten (über alle Fächer gemittelt), keine Verbesserung im Mathematik-Kurs
Grove/ Wasserman (2004)	1998 vs. 2002	1 Private Universität	Semesternoten (Undergraduate students)	Verbesserung um 0.11 Noten (über alle Fächer gemittelt)
Kezim et al. (2005)	1983-2003	1 Privates College (nur Business School),	Kursnoten (Undergraduate students)	Verbesserung um durchschnittlich 0.0075 Noten pro Jahr

noch Tabelle 3: Übersicht über zentrale Studien, die langfristige Notenverbesserungen im US-amerikanischen Hochschulsystem feststellen

Studie	Zeitraum	Untersuchungseinheit(en)	Untersuchte Noten	Zentraler deskriptiver Befund
Mulvenon/Ferritor (2005)*	1992/93-2003/04	University of Arkansas	Kursnoten (Undergraduate students/ Graduate students)	Verbesserung um 0.19 (Undergraduate)/0.05 (Graduate) Noten (über alle Fächer gemittelt)
Lowe et al. (2008)	1993 vs. 2000/01	Bundesweites sample (10028/10030 Studierende)	Abschlussnoten (Undergraduate students)	Verbesserung um 0.23 Noten (über alle Fächer gemittelt), fachspezifische Verbesserung
Achen/Courant (2009)*	1992-2008	College (Literature, Science, Arts) der Universität von Michigan, Auswertung nach Kurslevel	Kursnoten (Undergraduate Kurse)	Kurspezifische Verbesserung nach Einführungs- vs. Fortgeschrittenenlevel, auch Kurse ohne Verbesserung
Baird (2009)*	1980-2008	Alle Fakultäten an der Universität von California	Kursnoten (Undergraduate Kurse/ Graduate students)	Fachspezifische Verbesserung, auch Fächer ohne Verbesserungen
Bar et al. (2009)*	1990-2004	Cornell College	Kursnoten (Undergraduate Kurse)	Verbesserung um 0.18 Noten (über alle Fächer gemittelt)
Mathies/Webber (2009)	1985-2004	1 Öffentliche Universität	Notendurchschnitte (Undergraduate students)	Verbesserung um 0.69 Noten (über alle Fächer gemittelt)
Rush et al. (2009)*	1985-2006	College of Veterinary Medicine der Kansas State University, Auswertung nach Notenquantilen	Abschlussnoten (Undergraduate students)	Verbesserung um 0.416 Noten (über alle Studiengänge gemittelt), quantilspezifische Verbesserung
Wongsurawat (2009)	1995 vs. 2000 vs. 2007	48 Rechtswissenschaftliche Fakultäten	Abschlussnoten (Undergraduate students)	Verbesserungen im berechneten Inflationsindex (über alle Bewerber gemittelt) zwischen 1995 und 2000 sowie zwischen 2000 und 2007
Popov/Bernhardt (2010)	1960 vs. 1980 vs. 2000	20 Universitäten, Auswertung nach Platzierung im US News National Universities Ranking	Notendurchschnitte (Undergraduate students)	Verbesserung um 0.7613 (0.5933) Noten (über die Universitäten gemittelt) von 1960 bis 2000 an den besser (schlechter) platzierten Universitäten
Jewell/McPherson (2012*	1984-2005	Universität von North Texas, Auswertung nach Fakultäten	Kursnoten (Undergraduate Kurse)	Verbesserung um 0.375 Noten (über alle Fächer gemittelt), fakultätsspezifische Verbesserung
Rojstaczer/Healy (2012)	1940-2009	Über 200 Colleges und Universitäten (sample nicht konsistent im Zeitverlauf)	Semesternoten (Undergraduate students)	Anteil ‚A' ist von 15% auf 43% gestiegen, der Anteil ‚C' von 35% auf ca. 15% gesunken (über alle Hochschulen und Fächer gemittelt)
Jewell et al. (2013)*	1984-2005	Universität von North Texas, Auswertung nach Fakultäten	Kursnoten (Undergraduate Kurse)	Verbesserung um 0.363 Noten (über alle Fächer gemittelt), institutsspezifische Verbesserung, auch Institute ohne Verbesserung
gradeinflation.com	diverse	bis zu 88 Universitäten und Colleges (Stand 03/2016)	Notendurchschnitte/-verteilungen (Undergraduate students)	Unterschiedlich starke Verbesserung nach betrachtetem Zeitraum, fach- und hochschulspezifische Verbesserung, institutionenspezifische Verbesserung

Mit einem * markiert sind Studien, die eine Hochschule untersuchen, an der mindestens einer der Autor*innen angestellt ist.

6.3.3 Notenverbesserung vs. Noteninflation: Die Suche nach Ursachen

Die Masse an Befunden suggeriert eine klare Beweislage: Noteninflation scheint eine globale Entwicklung an US-amerikanischen Hochschulen darzustellen[43]. Zahlreiche Kommentare zu diesem Thema in den Medien stützen das Bild vom leistungsschwachen Studierenden, der im Gegensatz zu früher trotz mangelnder Begabung oder mangelnden Engagements mit Bestnoten überschüttet wird (etwa Mansfield 2001; Pedersen 1997; Staples 1998).

Es gibt allerdings mehrere Arten von Einwänden, die dabei helfen, dieses Bild etwas zurechtzurücken. Einmal gibt es neben den zahlreichen Befunden, die eine konstante Zunahme des Anteils guter Noten präsentieren, auch Studien, die von einer Abnahme der Geschwindigkeit, in der sich die Noten verbessern, berichten (Juola 1980; Bejar/Blew 1981). Außerdem existieren Studien, die keine Hinweise auf langfristig stabile Notenverbesserungen im Zeitverlauf und sehr gute Noten für die meisten Studierenden finden können (Adelman 1995 und 2008; Horn et al. 2002). Die Datenbasis dieser Studien unterscheidet sich dabei in wesentlicher Hinsicht gegenüber den meisten Studien, die eine Verbesserung von Noten im Zeitverlauf finden. Während diese Studien, die nach Juolas Survey durchgeführt wurden, häufig nur einzelne Colleges oder Universitäten oder aber, wenn hochschulübergreifend, dann Hochschulen mit vergleichbarem institutionellem Charakter, meist private und/oder Eliteuniversitäten, untersuchen[44], beruht Adelmans Analyse auf einer bundesweiten Stichprobe von Zeugnissen. Die Analysen von Horn et al. basieren sogar auf einer Befragung aller 16.5 Millionen US-amerikanischer Undergraduate Studierenden im Jahr 1999/2000. Diesen Zahlen nach hatten 14.5% der Studierenden bis zum Befragungszeitunkt hauptsächlich „A's", erhalten, 10.9% berichten von vorwiegend „A's and B's". Demgegenüber gaben 14.9% an, bisher hauptsächlich mit „C's", 14% vor allem mit „D's" und schlechter (24.6% v.a. „B's" 21.1% „B's and C's") bewertet worden zu sein. Studierende an privaten Universitäten erzielen im Durchschnitt bessere Noten als Studierende an öffentlichen Hochschulen.

---

[43] Anglin und Meng (2000) zeigen anhand von Erstsemesternoten von sieben kanadischen Universitäten, dass Notenverbesserungen im Zeitverlauf kein spezifisches Phänomen des US-amerikanischen Bildungssystems darstellen. Sie stellen fest, dass das Notenniveau an den Universitäten des Bundesstaats Ontario 1993/94 deutlich besser ist als 1973/74. Auch sie können für einige Fächer eine Abnahme der Streuung der Noten um den Mittelwert feststellen.

[44] Eine Ausnahme unter den Studien, die Notenverbesserungen im Zeitverlauf feststellen, stellt die Studie von Lowe et al. (2008) dar, die ein bundesweites sample nutzt. Allerdings vergleichen die Autor*innen nur die Noten zweier Zeitpunkte und schlussfolgern daraus einen Inflationsprozess. Da jedoch die Möglichkeit besteht, dass einer der Werte (oder sogar beide) einen Ausreißerwert darstellt, sollten die Ergebnisse eines derartigen Vorgehens eher skeptisch betrachtet werden. Levine und Cureton analysieren bundesweite Surveyergebnisse, allerdings verweist Kohn (2008) auf mangelnde Repräsentativität der Daten.
Auch Kuh und Hu (1999) nutzen Surveydaten, vergleichen mit den Mittelwerten aus zwei kurzen Zeitperioden (4 und 3 Jahre) ebenfalls nur zwei Werte miteinander, verringern dadurch aber immerhin das Ausreißerproblem. Das sample von Wongsurawat (2009), der Daten aus 48 Institutionen nutzt, umfasst nur Law Schools, deren Bewerber*innen in Hinblick auf ihren vorangegangenen akademischen Werdegang eine selektive Auswahl darstellen.

Die Ergebnisse von Adelman und Horn et al. legen nahe, dass das tatsächliche Ausmaß des Verbesserungsprozesses durch Verzerrungen bei der Auswahl der Untersuchungseinheiten überschätzt wird. So ist denkbar, dass erst das Bekanntwerden von Verbesserungsprozessen an Hochschulen die Autor*innen, die oftmals an derselben Hochschule angesiedelt sind, deren Noten sie untersuchen, dazu motiviert, die Notenentwicklung im Zeitverlauf zu analysieren (vgl. Tab.3: Die Autor*innen aller Studien, die an einer einzelnen Hochschule durchgeführt wurden und in denen diese Hochschule benannt wird, arbeiten auch dort). Auf diese Weise bliebe die Notenentwicklung an vielen Hochschulen ohne Verbesserung unberücksichtigt.

Doch auch dort, wo sich die Noten nachweislich verbessert haben, muss dies nicht unbedingt ein Problem darstellen. So existieren verschiedene Befunde, die nahelegen, dass eine Verbesserung von Noten nicht unbedingt mit einer Entwertung einhergehen muss. Eine Verschiebung der Notenverteilung zu mehr besseren Noten geht nicht unbedingt mit einem Reliabilitätsverlust (Millman et al. 1983) oder mit verringerten Differenzierungsmöglichkeiten zwischen Studierenden einher (Kuh/Hu 1999). Pattison et al. (2013) zeigen, dass die Noten an US-amerikanischen Colleges und Universitäten im Zeitverlauf nicht an Signalkraft verloren haben.

Schließlich lohnt sich ein weiterer Blick auf das empirische Material, dass die zuvor aufgeführten Studien enthalten. Bisher wurden lediglich deren Befunde hinsichtlich der Entwicklung auf Hochschulebene oder über mehrere Hochschulen hinweg aggregierter Noten vorgestellt. Wie im vorherigen Abschnitt jedoch aufgezeigt wurde, ist es zwingend notwendig bei der Analyse der Notenvergabe, sowohl hochschul- als auch fachspezifische Noten zu betrachten. Zahlreiche der aufgeführten Studien (Bar et al. 2009; Grove/Wasserman 2004; Juola 1976; Levine/Cureton 1998; Mathies/ Webber 2009; Mc Spirit/Jones 1999; Popov/Bernhardt 2010; Potter et al. 2001; Rogers 1983; Rojstaczer/ Healy 2012; Suslow 1976; Wongsurawat 2009) fassen die Noten der Studierenden aller Fächer (und teilweise mehrerer Hochschulen) zusammen, können möglicherweise unterschiedliche Entwicklungen innerhalb der einzelnen Organisationseinheiten also gar nicht erfassen.

Dass dies dringend notwendig ist, um dem Phänomen der Notenverbesserung gerecht zu werden, wird in den Studien deutlich, die entsprechende Differenzierungen vornehmen. So zeigen sich bei Achen und Courant (2009), Anglin und Meng (2000), Birnbaum (1977), Jewell und McPherson (2012), Kuh und Hu (1999), Lowe et al. (2008), Mullen (1995) und Sabot und Wakeman-Linn (1991) deutliche Unterschiede im Ausmaß, in dem sich die Noten in den untersuchten Fächer(gruppe)n, Fakultäten, Campus oder Instituten in den betrachteten Zeiträumen verbessert haben. Baird (2009), Cheong (2000), Compton und Metheny (2000), Jewell et al. (2013), Kolevzon (1981) sowie Prather et al. (1979) stellen zudem fest, dass der Trend zu besseren Notendurchschnitten an den von ihnen jeweils untersuchten Hochschulen nicht in allen Fächern wiederzufinden ist, Kurse in den Erziehungswissen-

schaften und in Soziologie weisen bei Erstgenannten sogar schlechter werdende Noten auf. Auch bei Anglin und Meng (2000) sind die Noten in Soziologie zum späteren Vergleichszeitpunkt schlechter als zuvor.

Diese Einschränkungen in Kombination mit der Vermutung, dass die Befunde, die für eine konstante Notenverbesserung im Zeitverlauf sprechen, möglicherweise ein verzerrtes Abbild der Gesamtpopulation darstellen, verdeutlichen, dass die kontinuierliche Verbesserung von Noten im Zeitverlauf nicht ohne weiteres als allumfassendes, einheitliches Phänomen akzeptiert werden sollte. Es existieren Unterschiede in der Geschwindigkeit, in der sich die Noten an einzelnen Fächern und Hochschulen verbessern und es gibt Ausnahmen vom allgemeinen Trend (möglicherweise viele noch unbekannt). Beide Relativierungsaspekte werden in Studien, die die Noten nur über mehrere Fächer oder Hochschulen hinweg betrachten, verschleiert.

Die Beweisführung zugunsten der Existenz von grade inflation, also der Entwertung von Noten im Zeitverlauf, weist zudem oftmals einen erheblichen Mangel auf. Sie ist, wenn überhaupt, üblicherweise als residuale Erklärung konzipiert. Bewusst wurde als gemeinsamer Befund der sich an Juolas Untersuchung anschließenden Studien stets nur die *Verbesserung* von Durchschnitts-, Kurs- oder Abschlussnoten herausgestellt. Denn bereits bei der als Ursprung der Forschung zum Phänomen Noteninflation angesehenen Studie Juolas findet sich eine genau genommen unzulässige Schlussfolgerung: Die festgestellte Verbesserung der Noten wird ohne jede Begründung als inflationärer Entwertungsprozess eingestuft.

Wie auch bereits für im Querschnitt festzustellende Unterschiede im Notenniveau muss jedoch auch für Unterschiede im Zeitverlauf beachtet werden, dass sie auf verschiedene Arten entstehen können. Verbesserte Notenniveaus im Zeitverlauf können ebenfalls durch leistungskonforme und leistungsunabhängige Faktoren erklärt werden. Theoretisch könnten bessere Leistungen der Studierenden (z.B. bedingt durch eine verbesserte Ausbildung vor dem Hochschuleintritt, durch gestiegenes Engagement oder: durch immer bessere Lehre, etwa bei Falkenberg 1996) als Erklärung für bessere Notenniveaus herangezogen werden.

Und auch wenn sich die Leistung der Studierenden als solche nicht verbessert haben sollte, muss dies nicht zwangsläufig bedeuten, dass Studierende im Zeitverlauf immer weniger für gute Noten leisten müssen. So könnten verbesserte Notenniveaus auch durch veränderte Zusammensetzungen der Studierenden (veränderte Geschlechterkomposition, veränderte soziale Zusammensetzung, veränderte Altersstruktur - siehe Birnbaum 1977) bedingt werden, durch die die Gesamtleistung zu steigen vermag. Ein Selektionsprozess, der auf der Existenz von unterschiedlichen Notenniveaus beruht, und dazu führt, dass Studierende zunehmend Kurse, Fächer oder auch Hochschulen mit bekanntermaßen besseren Noten wählen, könnte eine Verbesserung von Notenniveaus ebenfalls ohne eine Veränderung des Verhältnisses zwischen individueller Leistung und erhaltener Note erklären (ebd.; Sa-

bot/Wakeman-Linn 1991). Erst wenn solche kompositionellen Faktoren, wie auch alle anderen leistungskonformen Erklärungen ausgeschlossen werden können, sollte in Betracht gezogen werden, dass sinkende Bewertungsstandards die Ursache für verbesserte Notenniveaus darstellen.

Unter Berücksichtigung dieser Prämisse schrumpft die zunächst überwältigend erscheinende Evidenz der Studien. Die Arbeiten von Anglin und Meng (2000), Compton und Metheny (2000), Grove und Wasserman (2004), Juola (1976), Levine und Cureton (1998), Suslow (1976) und Wongsurawat (2009) können aus dieser Perspektive nicht als Beleg für eine Entwertung von Studienleistung betrachtet werden. Sie stellen verbesserte Notenniveaus fest und zeigen damit ein erklärungsbedürftiges Phänomen auf, dessen postulierte Ursache allerdings bestenfalls auf theoretischen Überlegungen beruht[45]. Achen und Courant (2009) sehen die differenzierte Darstellung von Kursnoten zwar als Beweis dafür an, dass die Fakultäten mit Hilfe der Noten versuchen, studentisches Verhalten, insbesondere die Kurswahl, zu steuern, zeigen allerdings keine statistischen Zusammenhänge auf. Cheong (2000) untersucht einen möglichen Zusammenhang zwischen dem Frauenanteil und der festgestellten Notenverbesserung, allerdings ohne ihn zu finden. Popov und Bernhardt (2010) nutzen die aufgezeigten Verbesserungen als Aufhänger für eine Modellierung des Inflationsverhaltens von Hochschulen, zielen aber nicht auf die Erklärung dieser konkreten Werte ab.

Die restlichen Studien weisen hier schon einen höheren Erklärungsgehalt auf. Sie überprüfen den Einfluss unterschiedlicher Faktoren auf die Notenentwicklung oder den Zusammenhang zwischen diesen Faktoren und der Notenhöhe im Allgemeinen[46]. Baird (2009) und Cluskey et al. (1997) versuchen, tatsächliche grade inflation vom gesamten Verbesserungsprozess zu isolieren. Sie kontrollieren die Eingangsleistung der Studierenden, die einen signifikanten Einfluss auf die Notenhöhe zeigt, nur die nach Kontrolle der Leistung bestehenden Verbesserungen werden als Evidenz für grade inflation betrachtet. Auch Mullen (1995) zeigt, dass eine zunehmende Eingangseignung zur Verbesserung der Noten beiträgt. Mulvenon und Ferritor (2005) und Kwon et al. (1997) können eine Verbesserung im beobachteten Zeitraum ebenfalls feststellen, Erstere außerdem schlechtere Noten bei unbefristet beschäftigten (tenure) Lehrenden, Letztere ein gestiegenes Durchschnittsalter der Studierenden aufzeigen, während Rojstaczer und Healy (2012) zwar einen starken Zusammenhang zwischen SAT-

[45] Koedel (2011) geht so weit, grade inflation aus Notenunterschieden zwischen Fächern zu nur einem einzigen Zeitpunkt abzuleiten. Er kann keine strukturellen Unterschiede für die besseren Noten von Lehramtsstudierenden zu einem Messzeitpunkt finden und schließt daraus, dass grade inflation im Lehramtsstudium für die unterschiedlichen Notenniveaus verantwortlich sein muss.

[46] Generell muss jedoch in Hinblick auf die in den vorgestellten Studien nachgewiesenen Zusammenhänge zwischen einzelnen unabhängigen Variablen und dem Notenniveau betont werden, dass diese Zusammenhänge alleine noch keine Verbesserungen im Notenniveau erklären können. Nur wenn sich das Verhältnis zwischen erklärender Variable und Notenhöhe im Zeitverlauf zugunsten des Einflusses in Richtung besserer Noten verändert oder dieser Einfluss im Zeitverlauf zunimmt, kann eine Verbesserung des Notenniveaus auch auf den Zusammenhang zwischen den beiden Größen zurückgeführt werden.

Scores und Noten aufzeigen, jedoch auch darauf verweisen, dass die SAT-Scores in ihrem sample im Zeitverlauf nicht ansteigen. Auch bei Potter et. al (2001) und Lowe et al. (2008) verbessert sich die Eingangseignung der Studierenden ihrer Untersuchungseinheiten im betrachteten Zeitraum nicht.

Baird weist zusätzlich noch auf deskriptive Hinweise auf einen Zusammenhang zwischen der Kursgröße und dem Notenniveau (bessere Noten in kleineren Kursen) hin, Lowe et al. führen chi^2 Tests durch, die keine signifikanten Änderungen in den demographischen Merkmalen und dem Fachwahlverhalten der Studierenden ausmachen können. Cluskey et al. und Potter et al. prüfen keine alternativen Erklärungen. Birnbaum (1977) überprüft die beobachtete Notenverbesserung auf Zusammenhänge mit der Entwicklung der Leistung im Studium, der Eingangseignung in Form von High School Abschlussnoten, des Geschlechterverhältnisses, der formalen Prüfungsbedingungen sowie der Fach- und Kurswahl. Seine Analysen zeigen dabei, dass veränderte Prüfungsbedingungen einen Einfluss auf das Notenniveau haben: Studierende nutzen Möglichkeiten zur Vermeidung von schlechten Noten erfolgreich. Geisinger (1979) stellt damit übereinstimmend fest, dass es einen mittleren positiven Zusammenhang zwischen Notenhöhe und der Abmeldungsquote in laufenden Kursen gibt, der durch die Einführung der Möglichkeit einer späten Abmeldung im laufenden Semester begünstigt wird: Können Studierende sich solange von Kursen abmelden, bis sich herauskristallisiert hat, in welchem Bereich ihre Note wohl liegen wird, ermöglicht es Ihnen taktische Abmeldungen, die vorher nicht möglich waren.

Kolevzon (1981) trennt Institute mit hoher Verbesserungsrate von Instituten mit geringer Verbesserungsrate und testet dann mögliche Einflussfaktoren auf statistisch signifikante Differenzen zwischen den Institutsgruppen. Er stellt dabei fest, dass Institute mit hoher Verbesserungsrate im betrachteten Zeitraum einen geringeren Zuwachs an männlichem gegenüber weiblichem Lehrpersonal und einen größeren Zuwachs an männlichen Studierenden verzeichnen. Das Lehrpersonal in den Instituten mit stärkerer Notenverbesserung nimmt zudem häufiger (24.3%) einen Zuwachs der Kursgrößen um mindestens 50% wahr als das Lehrpersonal in Instituten mit geringeren Verbesserungsraten (7.7%) und stimmt häufiger der Aussage, dass Aufgaben jenseits der Lehre Zeit kosten, die zur Benotung nötig wäre, zu (75.0% vs. 52.4%). Es befürwortet außerdem zu größeren Teilen die Möglichkeit zur nachträglichen Notenverbesserung für Studierende (64.6% vs. 43.6%) und setzt seltener standardisierte Prüfungsverfahren ein (30.2% vs. 42.9%, Unterschied bei $p<0.07$).

Rogers (1983) überprüft mit Hilfe zeitreihenanalytischer Verfahren den Einfluss der Studierendenleistung in Form von ACT-Scores, der Geschlechterkomposition, der Studierendenzahlen sowie des Verbraucherzeitindexes. Er findet lediglich einen schwachen Zusammenhang zwischen der Entwicklung der Studierendenzahlen und der Entwicklung der Noten, weist dabei selbst auf den explorativen Charakter seiner Studie und methodische Einschränkungen hin.

Mc Spirit und Jones kontrollieren die von ihnen festgestellten Verbesserungen in Abschlussnoten auf Leistung (über ACT-Scores) und Geschlecht und stellen für beide Variablen einen signifikanten Einfluss fest, nach dem höhere Testscores auch zu besseren Noten führen und Frauen bessere Noten erhalten als Männer. Den auch unter Berücksichtigung der Kontrollvariablen bestehenden signifikanten Einfluss des Eintrittsjahres in das Studium auf die Note deuten die Autor*innen als Nachweis für grade inflation. Der im Vergleich besonders hohe Einfluss des Eintrittsjahres für die Subgruppe der Studierenden mit niedrigen ACT-Scores sehen sie als Hinweis darauf, dass das Lehrpersonal gute Noten als Anreizinstrument verwendet, um diese Studierenden zum Lernen zu motivieren (ein Argument, das auch Zubrickas 2010 vertritt). Problematisch ist, dass Mc Spirit und Jones einfache OLS (Ordinary Least Squares)-Regressionen anwenden, um den Einfluss des Eintrittsjahres auf die Notenhöhe zu berechnen. Die Werte der Noten, die für den Zeitraum von 1983-1996 in die Rechnungen eingehen, stellen allerdings Zeitreihenwerte dar, was in den OLS-Regressionen mit hoher Wahrscheinlichkeit zum Problem autokorrelierter Residuen und damit zu verzerrten Schätzungen führt (Cochrane/Orcutt 1949; Stier 2001). Die Autor*innen geben keine Hinweise, ob sie die Daten auf Autokorrelation überprüft haben.

Kuh und Hu (1999) prüfen den Einfluss folgender unabhängiger Variablen auf selbstberichtete Noten, deren Entwicklung zwischen den Zeiträumen 1984-1987 und 1995-1997 sie untersuchen: Geschlecht, Ethnizität, sozioökonomischer Status (SES), Selektivitätsindex für US-amerikanische Colleges (Barron's Profile of American College, bildet die Aufnahmebedingungen der Hochschulen ab und dient damit als Indikator für die Eingangsbefähigung der Studierenden), institutioneller Charakter (privat vs. öffentlich), Fachsemester, die für das Lernen zuhause verwendete Menge an Zeit und drei weitere Items, die das Lernengagement abbilden. Die Zeitvariable behält ihren Einfluss auf die Notenhöhe auch unter Kontrolle dieser Variablen sowohl für die gesamte Stichprobe als auch in getrennten Analysen für einzelne Fächergruppen und Hochschulausrichtungen (gemäß Carnegie Foundation Kategorien).

Auch das Geschlecht, die Hautfarbe, der sozioökonomische Status, der Selektivitätswert, der institutionelle Typ der Einrichtung, das Fachsemester (zwei der drei Dummys) sowie alle leistungsabbildenden Variablen weisen über beide Zeiträume gemeinsam betrachtet einen signifikanten Einfluss auf die Notenhöhe auf: Weibliche und weiße Studierende sowie höhere Fachsemester berichten von besseren Noten. Privatuniversitäten vergeben bessere Noten als öffentliche Universitäten, auch ein höherer Selektivitätsrang begünstigt diese. Mehr Engagement führt ebenfalls zu besseren Noten, mit der Ausnahme, dass ein Anstieg der Zeit, die in der Bibliothek verbracht wird, mit schlechteren Noten einhergeht. Fach- und institutionenspezifische Auswertungen zeigen, dass diese Einflussfaktoren auf die Notenhöhe allerdings nicht allgemeingültig sind. In getrennten Berechnungen für die unterschiedlichen Ausrichtungen der Hochschulen weisen nur die Ethnizität, der institutionelle Charakter

und zwei der Leistungsvariablen in allen nach Carnegie Kategorien getrennten Universitäten einen signifikanten Einfluss auf. Die Trennung nach Fächergruppen zeigt einen fachübergreifenden Einfluss von Geschlecht, Ethnizität, institutionellem Charakter und allen vier Leistungsvariablen.

In die Regressionsanalysen aufgenommene Interaktionsterme, gebildet jeweils aus der Zeitvariable und den einzelnen unabhängigen Variablen, weisen darauf hin, dass sich der Einfluss des Geschlechts, des institutionellen Charakters, des Fachsemesters, des Engagements, der Eingangseignung und der Ethnizität auf die Noten im Zeitverlauf verändert hat. Allerdings zeigt sich dieser Effekt nicht für jede der Variablen auch an jedem Hochschultyp. Ob die Variablen signifikanten Einfluss aufweisen, hängt vom jeweiligen institutionellen Typ der Hochschule ab, nach dem die Analysen getrennt durchgeführt wurden. Getrennte Auswertungen nach Fächergruppen zeigen einen signifikanten Einfluss der Interaktion zwischen Zeit und Geschlecht, zwischen Zeit und sozioökonomischem Status sowie zwischen Zeit und institutionellem Charakter in den Geistes- sowie in den Sozialwissenschaften. In Ersteren hat sich zudem der Einfluss des Engagements, in Letzteren der der Eingangseignung und der Ethnizität im Zeitverlauf verändert. In den Mathematisch-Naturwissenschaftlichen Fächern weisen die Interaktionsterme in keinem Fall signifikante Werte auf, in den anwendungsorientierten Fächern ist dies bei der Eingangseignung, der Ethnizität und dem institutionellen Charakter der Fall.

Kezim et al. (2005) zeigen, dass ein Teil der von Ihnen festgestellten Verbesserung im Notenniveau auf den zunehmenden Anteil an eingesetzten Privatdozierenden zurückgeführt werden kann. Mathies und Webber (2009) kontrollieren die aufgezeigte Verbesserung in den Semesternoten auf ACT-Scores und High School Abschlussnoten sowie auf das Geschlecht, die Ethnizität und den Status als Stipendiat*in und als Transferstudierende*r. Höhere SAT-Scores und bessere High School Abschlüsse tragen dabei zur Erklärung der Verbesserung zwischen 1985 und 2004 bei. Auch das weibliche Geschlecht, eine weiße Hautfarbe und der Status als Stipendiat*in stehen sowohl über den gesamten Zeitraum hinweg sowie für 1985 und 2004 getrennt betrachtet im Zusammenhang mit sich verbessernden Noten. Der Einfluss des Geschlechts und des Stipendiat*innen -Status ist 2004 stärker, für die anderen Variablen schwächer, als 1985. Insgesamt weisen die berücksichtigten Variablen nur eine geringe Erklärungskraft auf. Rush et al. (2009) können in Korrelationsanalysen keine signifikanten Zusammenhänge zwischen Alter bzw. Geschlecht und Notenverbesserung finden, auch gestiegene Studiengebühren und die Einführung einer Wiederholungsmöglichkeit für Kurse können die Verbesserungen im Notenniveau nicht erklären.

Jewell und McPherson (2012) überprüfen den Einfluss von Geschlecht und Ethnizität der Lehrenden sowie der ACT-Scores der Studierenden, der Fachsemesterverteilung in den Kursen und der Kursgröße auf die Notengebung und stellen fest, dass mit zunehmender Kursgröße und einem steigenden Anteil niedriger Fachsemester schlechtere Noten vergeben werden. Während die Eingangseignung

der Studierenden und die Ethnizität der Lehrenden keinen Effekt auf die Notenhöhe haben (in Jewell et al. (2013) zeigt sich ein kurvilinearer Zusammenhang zwischen der Eingangseignung und den Noten, nach dem bessere Eignung bis zu einem bestimmten Punkt zu besseren, danach zu schlechteren Noten führt), geben weibliche Lehrende bessere Noten als männliche und weisen auch eine leicht höhere Inflationsrate bei der Notenvergabe auf. Der steigende Anteil an weiblichen Lehrenden kann in diesem Fall also einen Teil der festgestellten Verbesserung in den Kursnoten erklären. Zunehmende Institutsgröße, operationalisiert als Anzahl der Lehrenden, führt in Instituten ohne Doktorand*innen-programme zu besseren Noten, in Instituten mit solchen Programme zu schlechteren Noten in den Kursnoten der Undergraduate Studierenden (ebd.).

Sabot und Wakemann-Linn (1991) kontrollieren zwar nicht die Leistungsentwicklung der Studierenden, zeigen aber einen Zusammenhang zwischen der Kurswahl und der Notenhöhe auf, der den festgestellten Trend zu besseren Noten erklären könnte. Diese Erklärung wird auch von Prather et al. (1979) gestützt, die zeigen, dass die Durchschnittsnoten aller Studierenden einer Universität aufgrund veränderter Kurs- bzw. Fachpräferenzen besser werden, obwohl sich die vergebenen Noten im Zeitverlauf nur in 23 von 144 berücksichtigten Kursen tatsächlich verbessern. Je nach zugehörigem Fach können die Autor*innen zudem einen Einfluss von Alter (bessere Noten bei höherem Alter) und Geschlecht (bessere Noten für weibliche Studierende) sowie der Eingangseignung, der Höhe angerechneter Studienleistungen anderer Hochschulen und des Studienfortschritts der Studierenden (schlechtere Noten mit fortschreitendem Studium) auf die Höhe der Kursnoten nachweisen.

Bar et al. (2009) zeigen, dass ein Teil der Notenverbesserung an der Cornell Universität auf einen Politikwechsel der Universität hinsichtlich der Transparenz von Kursnoten zurückgeführt werden kann. Eine ursprünglich zur besseren Einordnung des Werts der Noten online gestellte Homepage mit den Medianwerten der Noten der einzelnen Kurse wird von den Studierenden genutzt, um herauszufinden, in welchen Kursen die besten Noten vergeben werden. Die Publikation der Ergebnisse hat damit den Selektionsprozess in Kurse mit üblicherweise guten Noten noch verstärkt anstatt ihn abzumildern. Auch Johnson (2003) kann in einer quasi-experimentellen Studie einen Einfluss der durchschnittlichen Kursnoten auf die Kurswahl der Studierenden nachweisen. Er stellt fest, dass Studierende sich bei Kenntnis der Kursnoten doppelt so häufig für den Kurs, in dem durchschnittlich ein „A“ vergeben wird als für den Kurs in dem die Durchschnittsnote bei „B“ liegt, entscheiden.

Tabelle 4: Überblick über die in Studien, die Notenverbesserung nachweisen, überprüften Einflussfaktoren

Studie	Zeitraum	Überprüfte Einflussfaktoren	Methode
Juola (1976)	1960-1973	--	--
Suslow (1976)	1960-1974	--	
Birnbaum (1977)	1968/69 vs. 1974/75	- Eingangseignung - Leistungsentwicklung	Regressionsanalyse
		- Geschlecht - Fachwahl - Kurswahl	Mittelwertvergleich
		- Formale Prüfungsbedingungen*	Simulation
Prather et al. (1979)	1970-1975	- Alter* - Geschlecht* - Minderheitenstatus - Veteranenstatus - Zeit seit dem High School Abschluss - Eingangseignung* - Studienfortschritt (Anzahl erreichter Credits)* - Studienleistung (Notendurchschnitt) - An anderen Hochschulen erbrachte Studienleistung (Notendurchschnitt und Credits)*	Regressionsanalyse
Kolevzon (1981)	1969/70 vs. 1975/76	- Beschäftigungsstatus der Lehrenden (Voll-/Teilzeit; Anteil Doktorand*innen/Assistenzprofessor*innen) - Geschlechterkomposition der Lehrenden* - Einstellungen der Lehrenden zum Verhältnis zwischen Lehrenden und Studierenden - Betreuungsrelation - Studierendenzahl - Studierendenstatus (Vollzeit- vs. Teilzeitstudium) - Geschlechterkomposition* - Einstellungen der Lehrenden zur Prüfungspraxis* - Art der eingesetzten Prüfungsverfahren*	Mittelwertvergleich
Rogers (1983)	1929-1981	- Eingangseignung - Geschlechterkomposition - Studierendenzahl* - Wirtschaftliche Entwicklung (Verbraucherpreisindex)	Regressionsanalyse; z.T. Verwendung gelagter Daten
Sabot/Wakeman-Linn (1991)	1962/63 vs. 1985/86	- Kurswahl*	Regressionsanalyse; Simulation
Mullen (1995)	1987-1992	- Eingangseignung*	Regressionsanalyse
Cluskey et al. (1997)	1980-1995	- Eingangseignung*	Regressionsanalyse
Kwon et al. (1997)	1983/84 vs. 1993/94	- Eingangseignung* - Alter*	Deskriptiver Vergleich
Levine/Cureton (1998)	1969 vs. 1976 vs. 1993	--	--
Kuh/Hu (1999)	1984-87 vs. 1995-97	- Geschlecht* - Ethnizität* - Sozioökonomischer Status* - Eingangseignung* - Institut. Status der Hochschule (Öffentlich vs. Privat)* - Fachsemester* - Zeit, die für das Lernen zuhause aufgewendet wird* - Lernengagement*	Regressionsanalyse
Mc Spirit/Jones (1999)	1983-1996	- Eingangseignung* - Geschlecht*	Regressionsanalyse
Anglin/Meng (2000)	1973/74 vs. 1993/94	--	--
Cheong (2000)	1987-1999	- Geschlecht	Deskriptiver Vergleich
Compton/ Metheny (2000)	1988-1997	--	--
Potter et al. (2001)	1969-1999	- Eingangseignung	Deskriptiver Vergleich
Grove/ Wasserman (2004)	1998 vs. 2002	--	--

noch Tabelle 4: Überblick über die in Studien, die Notenverbesserung nachweisen, überprüften Einflussfaktoren

Studie	Zeitraum	Überprüfte Einflussfaktoren	Methode
Kezim et al. (2005)	1983-2003	- Beschäftigungsstatus der Lehrenden (Befristet/Unbefristet/Privatdozierende)*	Varianzanalyse; Regressionsanalyse
Mulvenon/Ferritor (2005)	1992/93-2003/04	- Beschäftigungsstatus der Lehrenden (Befristet/Unbefristet)* - Eingangseignung*	Deskriptiver Vergleich
Lowe et al. (2008)	1993 vs. 2000/01	- Geschlecht - Ethnizität - Alter - Soziale Herkunft - Eingangseignung - Fachwahl	chi^2-Test
Achen/Courant (2009)	1992-2008	--	--
Baird (2009)	1980-2008	- Eingangseignung*	Regressionsanalyse
		- Kursgröße*	Grafische Analyse
Bar et al. (2009)	1990-2004	- Kurswahl*	Simulation
Mathies/Webber (2009)	1985-2004	- Geschlecht* - Ethnizität* - Eingangseignung* - Status als Transferstudierende*r - Status als Stipendiat*in*	Regressionsanalyse
Rush et al. (2009)	1985-2006	- Geschlecht - Alter	Korrelationsanalyse
		- Studiengebühren - Formale Prüfungsbedingungen	Deskriptiver Vergleich
Wongsurawat (2009)	1995 vs. 2000 vs. 2007	--	--
Popov/Bernhardt (2010)	1960 vs. 1980 vs. 2000	--	--
Jewell/McPherson (2012)	1984-2005	- Geschlecht der Lehrenden* - Ethnizität der Lehrenden - Eingangseignung - Kursgröße* - Fachsemesteranteile*	Regressionsanalyse
Rojstaczer/Healy (2012)	1940-2009	- Eingangseignung*	Regressionsanalyse
Jewell et al. (2013)	1984-2005	- Eingangseignung* - Anzahl Lehrende im Institut* - Kursgröße* - Fachsemesteranteile*	Regressionsanalyse
gradeinflation.com	diverse	--	--

Mit einem * markiert sind die Merkmale, die einen Zusammenhang mit der Notenhöhe aufweisen

Die empirischen Ergebnisse zur Wirkung von Einflussfaktoren auf die Notenentwicklung, die sich den aufgeführten Studien entnehmen lassen, weisen also nur in eingeschränktem Maße auf einen Entwertungsprozess in der Notengebung hin. Einige Studien weisen trotz des Anspruches, Notenentwicklungen im Zeitverlauf zu erklären, nur generelle Zusammenhänge zwischen Notenhöhe und bestimmten Merkmalen nach. Verändern sich die untersuchten Merkmale aber im Zeitverlauf gar nicht, sind diese Zusammenhänge auch nicht zur Erklärung temporaler Muster geeignet.

Die Studien, die tatsächlich Wirkungszusammenhänge im Zeitverlauf darstellen, weisen unterschiedliche Entwicklungen in den Rahmenbedingungen für Lehre und in den formalen Prüfungsbedingungen nach, die es ermöglichen, bessere Noten für eine bestimmte Leistung zu erhalten, teilweise auch in den Zusammensetzungen der Lehrenden. Diese Entwicklungen müssen als Ursachen für einen

tatsächlichen Inflationsprozess angesehen werden. Die empirischen Ergebnisse weisen jedoch darauf hin, dass jeder dieser Ursachen für sich genommen nur eine geringe Erklärungskraft zukommt.
Die in einigen Studien aufgedeckten Zusammenhänge zwischen der Notenhöhe und dem Alter, dem Geschlecht sowie der Eingangseignung der Studierenden weisen auf leistungskonforme Ursachen der festgestellten Notenverbesserungen hin. Hinsichtlich dieser Merkmale sind die Ergebnisse über alle Studien hinweg betrachtet jedoch alles andere als eindeutig, was zum Teil möglicherweise durch hochschulspezifische Merkmale der Studierendenzusammensetzung erklärt werden kann. Auch ein zunehmend taktisches Verhalten in der Kurswahl muss als leistungskonforme Ursache für Notenverbesserungen eingestuft werden. Im Aggregat ist bei einer Zunahme taktischen Verhaltens zwar eine bessere Leistung als früher möglich, ohne dass die Studierenden insgesamt mehr dafür tun müssen als die Prüflinge vor ihnen, die ihre Kurs- oder Fachwahl weniger nach Benotungskriterien ausgerichtet haben - auf der individuellen Ebene ändert sich das Verhältnis von Leistung und Note aber nicht. Wenn sich das Angebot an Prüfungen nicht erhöht (mehr Wahlmöglichkeiten), wird es nicht leichter, eine bestimmte Note zu bekommen - es wählen einfach nur mehr Prüflinge den leichteren Weg. Dies stellt eine Option dar, die auch Prüflingen zuvor offen stand. Damit handelt es sich bei der Zunahme taktischen Verhaltens wie auch bei der Veränderung der Zusammensetzung der Studierenden um einen kompositionellen Faktor leistungskonformer Verbesserung, der sich nur im Aggregat auf die Entwicklung des Notenniveaus auswirkt.

Neben empirischen Studien, die versuchen, die Existenz von grade inflation zu überprüfen, existieren sowohl in der US-amerikanischen als auch in der europäischen Hochschulforschung zahlreiche theoretische Erörterungen zu den Einflussfaktoren auf die Notenentwicklung sowie Untersuchungen, die den Zusammenhang zwischen solchen Einflussfaktoren und der Notenhöhe unabhängig von einer zeitlichen Entwicklung überprüfen. Außerdem liegen einige statistische Modellierungen vor, die die Entwicklung von grade inflation unter bestimmten Bedingungen simulieren.
Reviews der Forschung zum Phänomen grade inflation ordnen die vermuteten Einflussfaktoren in der Regel entlang mehrerer Dimensionen (etwa Boretz 2004; Pressman 2007; Winzer 2002). Neben der Betrachtung von Studierendenmerkmalen und -zusammensetzungen (z.B. bessere Leistungen im Zeitverlauf, mehr weibliche Studierende, die bessere Noten erzielen, mehr ältere Studierende, die ernsthafter studieren, mehr Studierende aus nicht-akademischen Haushalten, die höheres Engagement zeigen, mehr Stipendiat*innen, die ihre Stipendien durch Leistung sichern müssen - siehe etwa Adelman 1995; Birnbaum 1977; Potter et al. 2001; Winzer 2002) gehören zu diesen Dimensionen: Gesellschaftliche Ereignisse und Entwicklungen, institutionelle Lehr-, Lern- und Prüfungsbedingungen, monetäre Anreize sowie die Einstellungen von Lehrenden.

Die *veränderte Zusammensetzung von Studierenden* wird vor allem in Hinblick auf eine im Laufe der Zeit durchschnittlich gestiegene Eingangseignung diskutiert. Zwar lässt sich, wie oben aufgezeigt, ein Zusammenhang zwischen der Eingangseignung der Studierenden und deren Noten nachweisen. Allerdings spricht nur wenig dafür, dass sich die Eingangseignung im Zeitverlauf auch entsprechend der nachweisbaren Verbesserungen im Notenniveau gesteigert hat und sie somit erklären kann. Neben den bereits erwähnten Studien, die auch nach der Kontrolle von Leistungsindikatoren bestehende Notenverbesserungen nachweisen können, existieren gemischte Evidenzen hinsichtlich der Frage, ob sich die Leistungen von US-amerikanischen Studierenden generell verbessert haben. Während dies in einigen Studien verneint wird (etwa Breland 1976; Stone 1995), finden sich in anderen Daten Verbesserungen von Leistungsindikatoren im Zeitverlauf (etwa bei Mullen 1995 oder Mulvenon/Ferritor 2005).

Der Beginn einer Notenverbesserung wird in den USA häufig mit dem Vietnamkrieg, im Laufe dessen Studierende mit guten Noten vor dem Einzug in die Armee bewahrt werden sollten und mit der Gleichstellung von ethnischen Minderheiten, die zu guten Noten als Fördermittel geführt habe, in Verbindung gebracht (Birnbaum 1977). Diese Entwicklungen können theoretisch das Einsetzen eines Verbesserungsprozesses erklären, aber nicht seinen weiteren Verlauf. Geeigneter sind dazu etwa Annahmen, die auf kontinuierlichen Arbeitsmarktentwicklungen aufbauen. Wie die, dass eine zunehmende Polarisierung zwischen Arbeitsplätzen mit hohem und niedrigem Anforderungsprofil die Vergabe guter Noten fördert, da Lehrende ihren Studierenden nicht die Chance auf gute Jobs nehmen möchten (Yang/Yip 2003). Oder die Vermutung, dass ein zunehmender Konkurrenzkampf um wenige gute Stellen bei immer mehr geeigneten Absolvent*innen zu besseren Noten beiträgt (Pressman 2007). *Gesellschaftliche Ereignisse und Entwicklungen* dieser Art wirken nicht direkt auf die Prüfungsbedingungen, beeinflussen aber, wie bereits beschrieben, womöglich die Selektionsneigung der Prüfenden.

Die in der Kategorie der *institutionellen Lehr-, Lern- und Prüfungsbedingungen* vermutlich am häufigsten diskutierte potentielle Ursache für Notenverbesserungen ist die studentische Evaluation der Lehrleistung. Es wird argumentiert, dass Lehrende sich gute Bewertungen im Austausch gegen gute Noten quasi erkaufen (Correa 2001; Hu 2005; McKenzie 1975). Einige Autor*innen sehen diese These durch Befunde gestützt, die eine nennenswerte Korrelation zwischen studentischer Bewertung und erwarteter bzw. tatsächlich erhaltener Note oder auch der Differenz zwischen erwarteter Note und aktuellem Notendurchschnitt finden (etwa Eiszler 2002; Isely/Singh 2005; Krautman/Sander 1999; Weinberg et al. 2009, ein umfassender Überblick ist in Johnson 2003 zu finden). Johnson (2003) zeigt

in einem quasi-experimentellen Design, dass Lehrende ihre Evaluationsergebnisse durch die Vergabe besserer Noten aufbessern können.

Der häufig postulierte kausale Prozess der Aufbesserung von Evaluationsergebnissen durch bessere Noten wird jedoch nicht von allen Forschenden anerkannt (etwa Mitchell 1998). Unter anderem betonen Kritiker*innen neben methodischen Mängeln in vielen Studien zu diesem Zusammenhang, dass die positive Korrelation auch darauf beruhen kann, dass gute Noten durch gute Lehre zustande kommen und diese Lehre zu Recht gut evaluiert wird (etwa Marsh/Roche 2000; zusammenfassend Gump 2007). Doch selbst unter der Annahme, dass gute Evaluationen tatsächlich durch gute Noten erkauft werden, bleibt eine zentrale Frage im Hinblick auf die Erklärungskraft dieses Mechanismus als treibende Kraft von grade inflation: Warum sollte sich das Notenniveau *kontinuierlich* durch den Einfluss von Evaluationen verbessern? Um einen derartigen Zusammenhang sinnvoll begründen zu können, müsste die Bedeutung von Evaluationen für den Karriereverlauf in ähnlichem Maße und Tempo zugenommen haben wie die Notenverbesserung. Da die studentische Evaluation der Lehre (SET=Student Evaluation of Teaching) an den meisten US-amerikanischen Universitäten zwischen Ende der 1960er und Anfang der 1970er Jahre eingeführt wurde (Centra 1993), kann aber auch hier höchstens von zeitlich begrenztem Erklärungspotential ausgegangen werden. Um eine kontinuierliche Verbesserung erklären zu können, müsste die Bedeutung von Evaluationen für den Karriereverlauf in ähnlichem Maße zugenommen haben wie die Notenverbesserung.

Gegen einen hohen Erklärungsbeitrag von Evaluationen zur Notenverbesserung im Allgemeinen spricht die Einschätzung, dass sie bei Personalentscheidungen relativ zu anderen Faktoren nur eine untergeordnete Rolle spielen (Murray 2005) - eine Einschätzung, die im Allgemeinen jedoch eher die Ausnahme darstellt (vgl. etwa Schneider 2013). Vor allem bei Auf- oder Abwertungen von Evaluationsergebnissen - etwa bei der Einführung von Bonuszahlungen für gute Evaluationen (Mangan 2009) - ist es, unabhängig von einem langfristigen Zusammenhang, plausibel, kurzfristige Auswirkungen dieser Eingriffe auf die Notenhöhe zu erwarten.

Gestiegene Studierendenzahlen und damit einhergehend zunehmende Kursgrößen und Betreuungsrelationen werden ebenfalls als potentielle Ursachen für eine Veränderung des Notenniveaus betrachtet. Zunehmend schlechtere Rahmenbedingungen könnten die Lehrenden dazu motivieren, den Studierenden im Ausgleich immer bessere Noten zu geben (Müller-Benedict/Tsarouha 2011) - ein Argument, dass bereits zur Erklärung unterschiedlicher Notenniveaus im Querschnitt bemüht wurde.

Spätestens an dieser Stelle sollte deutlich werden, dass die möglichen Ursachen für Differenzen im Notenniveau zwischen verschiedenen Untersuchungseinheiten zu einem gegebenen Zeitpunkt zu einem Großteil auch als mögliche Ursachen für Differenzen im Notenniveau einer einzelnen Untersuchungseinheit im Zeitverlauf in Frage kommen. So auch ein steigendes Prüfungsaufkommen, das Prüfende eines Studiengangs dazu bringen könnte, zunehmend bessere Noten zu vergeben, um die

Zeit, die sie für Forschung und Verwaltung benötigen, nicht für zunehmende Verhandlungen von Beschwerden über schlechte Noten nutzen zu müssen (Franz 2010; Yorke 2008). Dieser Mechanismus werde im Zeitverlauf besonders durch eine zunehmende Anzahl von leistungsschwachen Studierenden angetrieben, so Stone (1995). Eine ähnliche Argumentation verfolgt, wer behauptet, dass Studierende zunehmend prüfungsorientiert lernen und Lehrende ihnen dies ermöglichen, indem sie Prüfungsinhalte im Voraus immer genauer bekanntgeben (vgl. Winzer 2002). Ein derartiger Mechanismus könnte Teil eines umfassenderen Handels zwischen Studierenden und Lehrenden sein, den Kuh als „disengagement compact" (Kuh 2003:28) bezeichnet: Lehrende erwarten weniger von ihren Studierenden und vergeben bessere Noten, die Studierenden ihrerseits erwarten weniger Lehrleistung und damit einhergehenden Zeitaufwand von den Lehrenden, den die in den zunehmenden Aufwand für Forschung und Verwaltung umlenken.

Schließlich werden formale Regelungen der Prüfung, etwa Optionen zur Vermeidung von schlechten Noten, durch einen reinen Nachweis des Bestehens statt einer Note (Birnbaum 1977) , durch Freiversuche (Lanning/Perkins 1995) oder durch vereinfachte Rücktritte von Prüfungen (Stone 1995), wie auch gestiegene Wahlmöglichkeiten (Falkenberg 1996) und Bewertungsverfahren mit Notenverbesserungen in Verbindung gebracht. Für Letztere wird vor allem ein Wechsel von der sozialen zur sachlichen Bezugsnorm hervorgehoben, der die Differenzierung zwischen den Studierenden abschwächt und so mehr gute Noten trotz unveränderter Leistungen ermöglicht (vgl. Olsen 1997; Rojstaczer/Healy 2012; Yorke 2011). Außerdem wird der Art der Notendifferenzierung, also ob ganze oder differenzierte Noten vergeben werden, ein Einfluss auf die Note unterstellt: Beim Übergang von ganzen zu differenzierten Noten, entstehe eine Art Substitutionseffekt, durch den einer abgeschwächten Bestnote (also eine 1.3, 1 minus o.ä.) die gleiche Bedeutung zukomme wie zuvor einer ganzen Note schlechter (Suslow 1976).

Werden *monetäre Anreize* als Grund für die Verbesserung von Noten angeführt, wird sowohl ganzen Hochschulen, damit implizit als einheitliche Akteurinnen betrachtet, als auch einzelnen Instituten unterstellt, die Notenverbesserung bewusst zu fördern, um damit Outputindikatoren für Fördermittel, etwa Absolvent*innenzahlen, positiv zu beeinflussen und Wettbewerbsvorteile zu erhalten (De Paola 2008; Oleinik 2009; Stone 1995; Warning/Welzel 2005). Ein Interesse an möglichst hoher öffentlicher Reputation, die sich durch die Ausbildungsleistung anzeigen lässt, könnte hier ebenfalls einen Anreiz leisten (Yang/Yip, 2003). Lawler (2001) weist darauf hin, dass Reputationsunterschiede zwischen Hochschulen die Lehrenden an statusniedrigeren Einrichtungen zusätzlich dazu motivieren könnten, gute Noten zu vergeben, damit ihre Studierenden auf dem Arbeitsmarkt überhaupt neben den Studierenden aus Eliteuniversitäten bestehen können.

Wie die empirischen Befunde zeigen, entwickeln sich die Noten allerdings auch in den USA studiengangspezifisch in unterschiedlichem Tempo und nicht an jeder Einrichtung in jedem Studiengang gleichermaßen zum Besseren (etwa Kuh/Hu 1999; Sabot/Wakeman-Linn 1991). Vor dem Hintergrund, dass Hochschulen vor allem im letzten Jahrhundert überwiegend als lose gekoppelte Organisationen (Weick 1976) ohne effiziente zentrale Kontrollfunktionen betrachtet werden, sprechen hochschulinterne, fachspezifische Entwicklungen in der Notenvergabe eher gegen die Auffassung, Hochschulen würden eine Noteninflation systematisch fördern, indem sie als einheitliche Akteurinnen auftretend, Verbesserungen taktisch einsetzen.

Dass der Wettbewerb um finanzielle Mittel innerhalb der Hochschulen Notenverbesserungen produziert, ist aufgrund theoretischer Überlegungen wie auch empirischer Indizien eher haltbar. So werden in kleinen Instituten und solchen mit niedrigen Betreuungsrelationen die besseren Noten vergeben (Baird 2009; Dickson 1984). Den kleineren Hochschulen und Instituten wird ein Interesse an der Vergabe immer besserer Noten nachgesagt, um die Teilnehmer*innenzahlen zu halten oder zu erhöhen und im akademischen Betrieb konkurrenzfähig zu bleiben (etwa Staples 1998). Auch auf das konkrete Interesse der individuellen Lehrperson, vor allem befristet angestellter Lehrkräfte, ihren Arbeitsplatz nicht zu verlieren, wird hingewiesen (Dickson 1984; Stone 1995).

Diese Annahme wird durch Befunde gestützt, nach denen Privatdozierende und andere befristete Lehrkräfte bessere Noten als ordentliche Professor*innen vergeben (Barth et al. 2009; Moore/Trahan 1998; Sonner 2000; nur für Privatdozierende, nicht für befristete Lehrkräfte: Gohmann/McCrickard (2001); Kezim et al. 2005), was allerdings auch mit der These, dafür sei mangelnde Erfahrung in der Notengebung verantwortlich, kommentiert wird (etwa Foster/Foster 1998). Die Analysen von Gohmann und McCrickard sprechen für diese Einschätzung: Sie zeigen, dass Lehrende, die sich im Karriereverlauf auf die Entscheidung über eine Festanstellung zu bewegen zu diesem Zeitpunkt keine besseren Noten geben als sonst. Im Gegenteil weisen ihre Daten darauf hin, dass Lehrende mit zunehmender Dauer seit dem Zeitpunkt ihrer Festanstellung auch zunehmend bessere Noten vergeben. Dass befristet Angestellte im Durchschnitt dennoch bessere Noten vergeben als Festangestellte, erklären die Autor*innen damit, dass die mangelnde Erfahrung der Prüfenden sie im Zweifelsfalls dazu motiviert, bessere Noten zu vergeben, um Studierende nicht unberechtigterweise zu benachteiligen.

Wieder aus einer allgemeineren Perspektive wird das wechselnde Selbstverständnis von Hochschulen hin zu Dienstleisterinnen, die Studierende als Kunden betrachten, die für ihre Ausbildung zahlen und eine entsprechende Leistung dafür erhalten sollen, in den Fokus genommen. Die gewünschte Leistung stellt dabei nicht mehr primär Bildung dar, sondern ein auf dem Arbeitsmarkt optimal verwertbares Zertifikat (vgl. Kirp 2003; Rojstaczer/Healy 2012; Rosovsky/Hartley 2002).

Die *Einstellungen von Lehrenden* werden vor allem hinsichtlich möglicher Einflüsse auf ihre Selektionsneigung diskutiert. Differenzen in den Notenniveaus zwischen Fächern oder Hochschulen könnten Lehrenden, deren Noten vergleichsweise schlecht sind, einen Anreiz bieten, ihre Noten an bessere Niveaus anzupassen, um ihre Studierenden nicht zu benachteiligen (Achen/Courant 2009; Brighouse 2008). Gelegentlich wird eine Wende zu einem egalitären Lehr- und Lehrklima, möglicherweise auf zunehmend liberalen bzw. progressiven Einstellungen unter Lehrenden beruhend (vgl. Bar/Zussman 2012; Lawler 2001), postuliert, als dessen Folge eine verringerte Orientierung der Bewertung an fachlichen Kompetenzen zugunsten von Sympathien vermutet wird (Winzer 2002). Solche Einstellungen wirken theoretisch nicht nur auf die Notenvergabe, wenn sie sich, beeinflusst durch gesellschaftliche Ereignisse und Entwicklungen, bei Anteilen aktiver Prüfer*innen verändern, sondern auch bei Wechseln des Lehrpersonals, bei Veränderungen der Geschlechterkomposition, möglicherweise sogar bei ganzen Generationswechseln von Lehrenden (vgl. Birnbaum 1977; Kolevzon 1981).
Entsprechend muss neben einer veränderten Zusammensetzung von Studierenden auch eine veränderte Zusammensetzung von Lehrenden als Einflussquelle auf die Notenentwicklung in Betracht gezogen werden. Auch zunehmende Unterstützung bei der Stipendienbewerbung (Hu 2005) und ein Wandel des wahrgenommenen Bildungsauftrags weg von der Selektion hin zum Aufbau eines gesunden Selbstvertrauens von Studierenden werden für einen Trend zu besseren Noten verantwortlich gemacht (Lanning/Perkins 1995; Mansfield 2001).
Dass Lehrkräfte im Zeitverlauf zunehmend selbst gute Noten im Studium erhalten haben und deswegen auch als Prüfende gut benoten (Winzer 2002) klingt wenig plausibel, dürfte zukünftiges Lehrpersonal während des Studiums in der Regel doch wohl unabhängig vom durchschnittlichen Notenniveau zu den besten Studierenden gehören und vorwiegend beste Noten im Studium erzielen. Anpassungsdruck innerhalb der Institute könnte aber ein Grund sein, weshalb die Noten nicht schlechter werden (Koedel 2011). Schließlich wird die Vergabe zunehmend guter Noten sogar als Ausgleichshandlung für geringe Forschungsleistung betrachtet (Pressman 2007), durch die Lehrende ihr Selbstwertgefühl aufbessern. Um dieses Argument halten zu können, müsste jedoch auch die Zahl derer, deren Forschungsleistung vergleichsweise gering ausfällt kontinuierlich ansteigen.

Der Überblick über die in der (bisher fast ausschließlich auf das US-amerikanische Hochschulsystem bezogenen) Literatur diskutierten Einflussfaktoren auf die Entwicklung des Notenniveaus offenbart vor allem eins: Es existiert eine Vielzahl von Erklärungsansätzen, denen theoretisch Einfluss auf die Notenvergabe im Zeitverlauf zugesprochen wird. Die empirischen Evidenzen für die tatsächliche Wirkung der vermuteten Ursachen sind trotz der jahrzehntelangen Aufmerksamkeit, die das Thema grade inflation in den USA auf sich gezogen hat, allerdings begrenzt.

Die vorhandenen Ergebnisse weisen darauf hin, dass sowohl leistungsunabhängige (Veränderungen in den Rahmenbedingungen für Lehre/in den formalen Prüfungsbedingungen/in der Zusammensetzung der Lehrenden) als auch leistungskonforme (verstärkte Selektion in Kurse mit guten Noten) Gründe für die Verbesserung von Notenniveaus existieren.

Dass sich das Notenniveau an US-amerikanischen Hochschulen tatsächlich immer in eine Richtung, nämlich in die Richtung besserer Noten bewegt, ist empirisch allerdings keineswegs so deutlich abgesichert, wie es in der Diskussion zum Phänomen grade inflation häufig suggeriert wird. Es gibt Belege für die Verbesserung von Noten im Zeitverlauf und einzelne Hinweise auf eine sinkende Streuung der Noten um ihre Mittelwerte. Dass sich die Noten über mehrere Untersuchungseinheiten gemittelt, seien dies Fächer oder Hochschulen, verbessern, ist eindeutig belegt. Es ist allerdings unklar, ob sich tatsächlich in allen Fächern und an allen Hochschulen des US-amerikanischen Hochschulsektors die Noten stetig verbessert haben bzw. verbessern und wie stark diese Verbesserungen wirklich ausfallen, wenn sie nachweisbar sind.
Für diese Unklarheit verantwortlich sind vor allem zwei Faktoren: Eine Vielzahl verschiedener Operationalisierungen der abhängigen Variable, also eine ungleiche Auswahl der betrachteten Noten und eine Vielzahl von unterschiedlich gewählten Untersuchungseinheiten, die teilweise ganze Hochschulmittelwerte, über alle Fächer hinweg konstruiert, hervorbringt.
Die Ergebnisse der Studien, die die Entwicklungen an mehreren Hochschulen oder an einer Hochschule in mehreren Fächern vergleichen, zeigen, dass nicht alle Fächer an allen Hochschulen langfristige Notenverbesserungen aufweisen. Und auch die Untersuchungseinheiten, für die Verbesserungen nachgewiesen werden können, unterscheiden sich zum Teil in der Geschwindigkeit bzw. im Ausmaß, in dem die Verbesserungen sich vollziehen. Eine differenzierte Analyse der Ursachen von Verbesserungsprozessen muss berücksichtigen, auf welcher Wirkungsebene die potentiellen Einflussfaktoren anzusiedeln sind. Allgemeingültige Einflussfaktoren müssten ihre Wirkung auf das Notenniveau an jeder Hochschule und in jedem Fach entfalten, fach- und hochschulspezifische Ursachen können Unterschiede in den Entwicklungspfaden zwischen einzelnen Fächern und/oder Hochschulen erklären.

Die empirische Evidenzen, die dafür sprechen, dass es sich bei den nachgewiesenen Verbesserungen um eine tatsächliche Entwertung der Noten handelt, sind im Vergleich zu dem Eindruck, der in zahlreichen Zusammenfassungen zum Thema grade inflation vermittelt wird (etwa Johnson 2003; Winzer 2002) relativ gering. Diese Diskrepanz entsteht vor allem durch einen allzu sorglosen Umgang mit der Diagnose grade inflation, beruhend auf einer mangelnden Differenzierung zwischen einem tatsächlichen Inflationsprozess und leistungskompatiblen Notenverbesserungen (mit einigen Ausnahmen, etwa Birnbaum 1977; Hu 2005). Umgekehrt existieren ebenso wenige, wenn nicht sogar noch weni-

ger Evidenzen, die es erlauben, die festgestellten Verbesserungsprozesse als vorwiegend leistungskonform einzustufen.

Als empirisch gesichert kann aufgrund der existierenden Studien lediglich angesehen werden, dass es in einigen Fächern einen Trend zu besseren Noten gibt. Von wie vielen Hochschulen dieser Trend in den jeweiligen Fächern getragen wird und ob er eher linear auf die Bestwerte hinzu oder etwa in Zyklen verläuft (Kolevzon 1981) kann anhand des vorliegenden empirischen Materials nicht abschließend geklärt werden. Auch die Frage, ob und in welchem Maße die festgestellten Verbesserungen an US-amerikanischen Universitäten und Colleges tatsächlich einen Inflations-, also einen Entwertungsprozess darstellen und auf geringere Bewertungsstandards zurückzuführen sind, muss aufgrund mangelnder empirischer Evidenz an dieser Stelle offen bleiben. Um es mit Kohns Worten zu beschreiben: “No one has ever demonstrated that students today get As for the same work that used to receive Bs or Cs. We simply do not have the data to support such a claim” (Kohn 2008:3).

Einiges spricht dafür, dass die individuelle Leistung der Studierenden sich im Zeitverlauf zwar nicht verbessert hat, Änderungen in den Rahmenbedingungen für Lehre, in den formalen Prüfungsbedingungen und in den Zusammensetzungen der Lehrenden sowie verstärkte Selektionsprozesse der Studierenden in Kurse bzw. Fächer mit guten Noten zum Teil für den festgestellten Verbesserungstrend verantwortlich sind, ihn aber nicht vollständig erklären können.

In Zusammenhang mit entsprechenden Befunden wird außerdem schnell deutlich, dass Verbesserungsprozesse im Notenniveau nicht unabhängig von Unterschieden in der Höhe des Notenniveaus zwischen Fächern und Hochschulen ablaufen. Derartige, im Querschnitt nachzuweisende Unterschiede im Notenniveau können Veränderungen im Zeitverlauf mehr oder weniger stark begünstigen. Bereits die reine Existenz dieser Unterschiede führt Prüfende in Fächern, Abschlussarten oder Hochschulen mit vergleichsweise schlechten Notenniveaus möglicherweise dazu, bessere Noten zu vergeben, um ihre Prüflinge im Vergleich mit besser bewerteten Studierenden nicht zu benachteiligen.

Höchstens am Rande werden bei den Versuchen, Notenverbesserungen im Zeitverlauf zu erklären, Überlegungen darüber angestellt, welches Wirkungspotential die jeweils vermuteten Ursachen tatsächlich im Zeitverlauf aufweisen. Häufig reicht den Autor*innen ein irgendwie gearteter Zusammenhang zwischen einer Variable und dem Notenniveau um die mit guten Noten korrelierende Ausprägung dieser Variable als ursächlich für stetige Verbesserungen in den Noten einzustufen, ohne dabei näher zu beleuchten, ob es sich um einmalig auftretende Einflüsse (mit kurz-, mittel- oder langfristigen Folgen) oder langfristig existente, die in Abhängigkeit ihrer eigenen Ausprägung auf das Notenniveau wirken, handelt. Tabelle 5 fasst die möglichen Ursachen für im Zeitverlauf verbesserte Notenniveaus, parallel zur Zusammenfassung der potentiellen Ursachen für unterschiedliche Notenniveaus, in Abhängigkeit ihrer Anschlussstellen im Notengebungsprozess in übergeordnete Ursachenkategorien zusammen und unterteilt sie zusätzlich gemäß ihrer Wirkungsart. Der Punkt *Gesell-*

schaftliche Ereignisse und Entwicklungen ist dabei auf einmalig wirkende Ereignisse (z.B. der Vietnamkrieg) und kontinuierlich präsente Entwicklungen (z.B. die technologische Entwicklung) aufgeteilt.

Tabelle 5: Mögliche Ursachen für im Zeitverlauf verbesserte Notenniveaus (Kategorien)

	Leistungskonform Verbessertes Leistungsvermögen durch:	**Leistungsunabhängig** Gesunkene Bewertungsstandards durch:
Einmalig wirkend		**Änderungen im Prüfungsprozess** - Änderungen formaler Prüfungsbedingungen - Änderungen in den eingesetzten Prüfungsverfahren - zunehmende Standardisierung von Prüfungen **Änderungen im Bewertungsprozess** - veränderte Bezugsnormenorientierung **Änderungen im Selektionsklima** - gesellschaftliche Ereignisse
Langfristig wirkend	**Individuell zunehmende Leistung** - zunehmende Eingangseignung - zunehmende Lehrqualität **Im Aggregat zunehmende Leistung** - zunehmende Selbstselektion (in traditionell besser benotende Fächer/Kurse/Prüfungen) - veränderte Zusammensetzung der Studierenden (z.B. Geschlecht, Alter, Ethnizität, soz. Herkunft, Stipendiat*in)	**Änderungen im Bewertungsprozess** - veränderte Häufung von Wahrnehmungsfehlern **Änderungen im Selektionsklima** - gesellschaftliche Entwicklungen - gestiegene Prüfungsbelastung - veränderte Rahmenbedingungen für Lehre - veränderte Zusammensetzung der Lehrenden (z.B. pol./ päd. Einstellungen, Geschlecht Generationenwechsel) - veränderte Arbeitsmarktchancen der Prüflinge - veränderte Finanzierungsstrukturen - unterschiedliche Notenniveaus

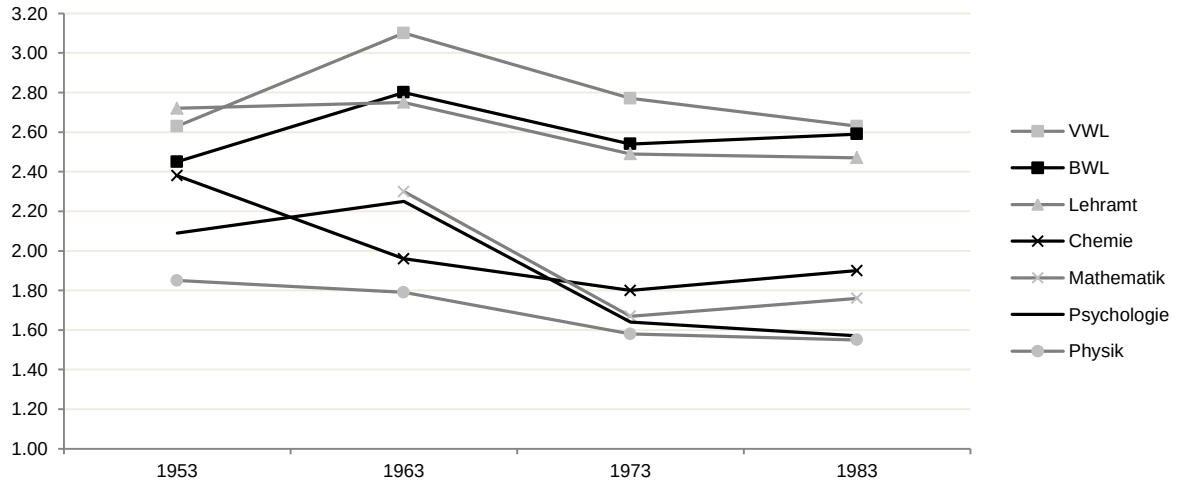
6.3.4 Empirische Befunde zur Notenverbesserung in Deutschland

Während die Existenz von unterschiedlich hohen Notenniveaus in Deutschland zumindest für die letzten Jahrzehnte genauso unzweifelhaft nachweisbar ist wie in den USA, ist die langfristige Entwicklung der Noten aus empirischer Sicht hierzulande vernachlässigt worden. Zwar erkennen bereits Hitpass und Trosien (1987) in ihren Daten eine langfristige Verbesserung der Abschlussnoten an deutschen Hochschulen auf Fachebene, ein genauer Blick auf die Notenentwicklung über die vier Zeitpunkte offenbart jedoch, dass sich über diese Interpretation streiten lässt.

Zwar lässt sich für vier der sechs genauer dargestellten Diplomstudiengänge (Physik, Chemie, Mathematik, Psychologie) sowie für die Lehramtsstudiengänge feststellen, dass die Noten 1983 besser sind als zum ersten Messzeitpunkt 1953. Lediglich in einem der sechs, nämlich in Physik, ist aber tatsächlich eine *kontinuierliche* Verbesserung vom ersten Erhebungszeitpunkt 1953 bis zum vierten Erhebungszeitpunkt 1983 festzustellen. In VWL und Psychologie sinkt das Notenniveau von 1963 an, ebenso in den Lehramtsstudiengängen. Der Abwärtstrend Letzterer wird vor allem durch die Lehramtsfächer Biologie und Physik getragen, in denen die Noten ab 1963 kontinuierlich sinken, während sich in Deutsch und Geschichte nach einer Verbesserung in 1973 wieder eine Verschlechterung in 1983 feststellen lässt. Bei den Lehramtsprüflingen in Englisch und Mathematik lässt sich gar kein

Abwärtstrend feststellen[47]. In BWL sind die Noten 1963 und 1983 im Durchschnitt schlechter als zu den jeweils vorherigen Messzeitpunkten 1953 und 1973. In Mathematik und Chemie sinkt der Durchschnitt von 1953 bis 1973, steigt dann jedoch wieder an. Die Spannweite zwischen den Studiengängen steigt, bedingt vor allem durch die Verschlechterung in VWL, 1963 gegenüber 1953 um knapp eine halbe Note an, sinkt an den beiden letzten Messzeitpunkten dann wieder um ca. 0.1 Noten ab. Der lange Zeitraum zwischen den Messzeitpunkten, immerhin jeweils 10 Jahre, lässt allerdings offen, wie sich das Notenniveau zwischen den Datenpunkten entwickelt hat. Zudem birgt die Betrachtung einzelner Messzeitpunkte das Risiko, dass einer (oder mehrere) dieser Datenpunkte einen Ausreißerwert darstellt, der den tatsächlichen Trendverlauf nicht widerspiegelt.

Abbildung 20: Durchschnittliche Abschlussnoten in sechs Diplomstudiengängen und im Lehramt (1.SE) - vier Zeitpunkte



Quelle: Hitpass/Trosien 1987, eigene Darstellung

Abbildung 21: Durchschnittliche Abschlussnoten in sechs Fächern mit Abschluss Lehramt (1.SE) - vier Zeitpunkte

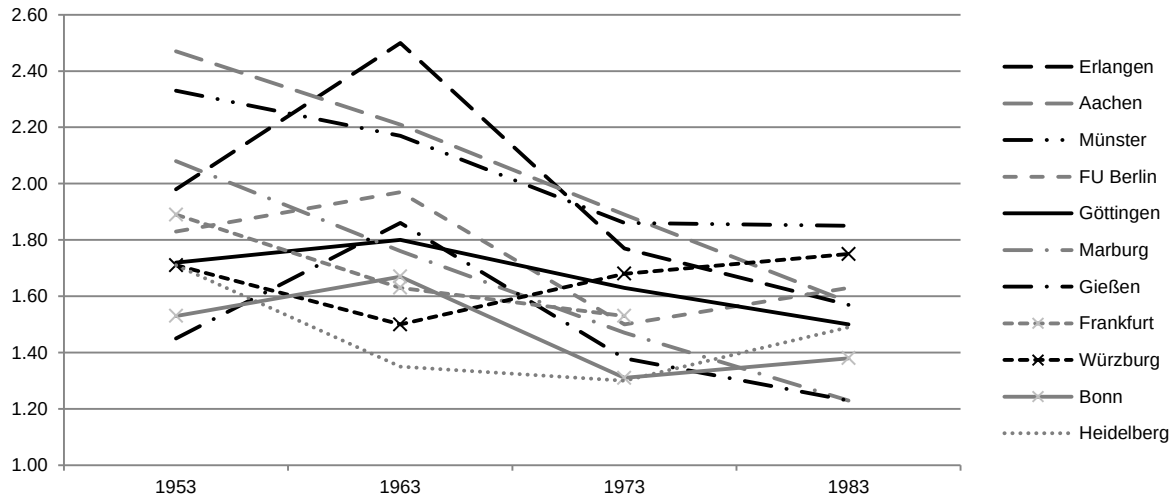
Quelle: Hitpass/Trosien 1987, eigene Berechnungen

[47] Die Verbesserung von Abschlussnoten im Lehramt über mehrere Fächer gemittelt, zeigt sich auch für das Lehramt an Volksschulen - für einen kürzeren Zeitraum (15 Semester), aber anhand kürzerer Messabstände nachgezeichnet (Ziegenspeck 1999). Aus der zugehörigen Grafik lässt sich eine Verbesserung von knapp einer halben Note im betrachteten Zeitraum feststellen. Die fachlichen Unterschiede in den Durchschnittsnoten, die im Querschnitt aufzeigt werden, bleiben bei der Längsschnittbetrachtung allerdings unberücksichtigt.

Bei einem Blick auf die Entwicklung der Notenniveaus an den einzelnen Hochschulen wird deutlich, dass auch an deutschen Hochschulen nicht nur unterschiedliche Notenniveaus im gleichen Studiengang zwischen Hochschulen bestehen, sondern auch keine parallele Notenentwicklung stattzufinden scheint. Auch wenn sich die Notenniveaus einiger Hochschulen in ihrer Entwicklung ähneln, gibt es immer Ausnahmen von einem möglichen Gesamttrend.

In Physik zeigen Marburg und Aachen einen durchgehenden Abwärtstrend, auch in Münster und Frankfurt (keine Daten für 1983) verbessert sich der Durchschnitt von jedem Messzeitpunkt zum nächsten. In Göttingen, Gießen und Erlangen sinkt das Notenniveau von 1963 an. In Berlin und Bonn steigt es 1963 und 1983 gegenüber 1953 bzw. 1973 an, in Heidelberg sinkt es bis 1973 und steigt dann wieder. Die deutliche Ausnahme vom allgemeinen Abwärtstrend stellt in Physik die Universität Würzburg dar. Hier ist das Notenniveau lediglich 1963 besser als 1953, bis 1983 wird es dann aber deutlich schlechter.

Abbildung 22: Abschlussnoten im Diplomstudiengang Physik an einzelnen Hochschulen - vier Zeitpunkte



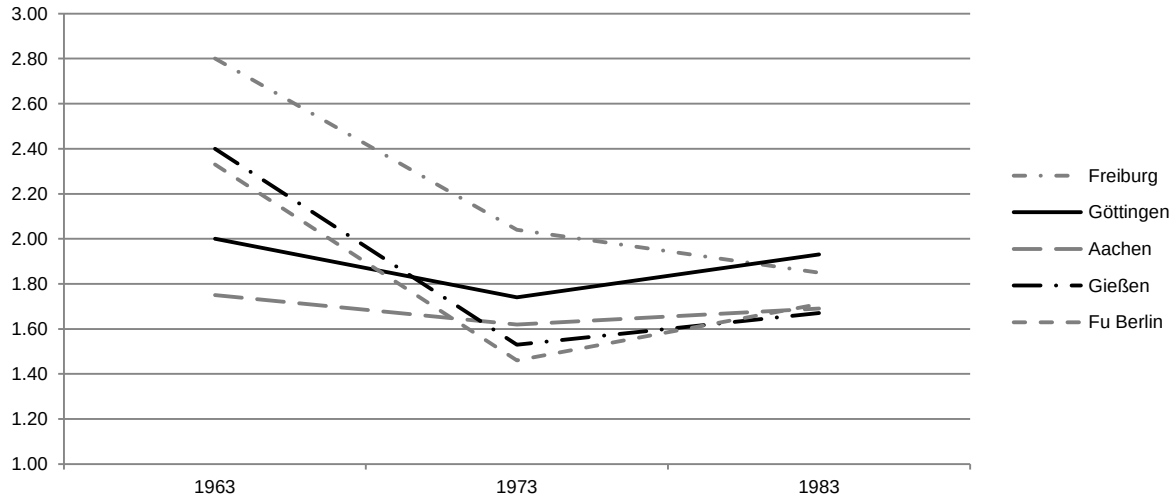
Quelle: Hitpass/Trosien 1987, eigene Darstellung

Abbildung 23: Abschlussnoten im Diplomstudiengang Chemie an einzelnen Hochschulen - vier Zeitpunkte

Quelle: Hitpass/Trosien 1987, eigene Darstellung

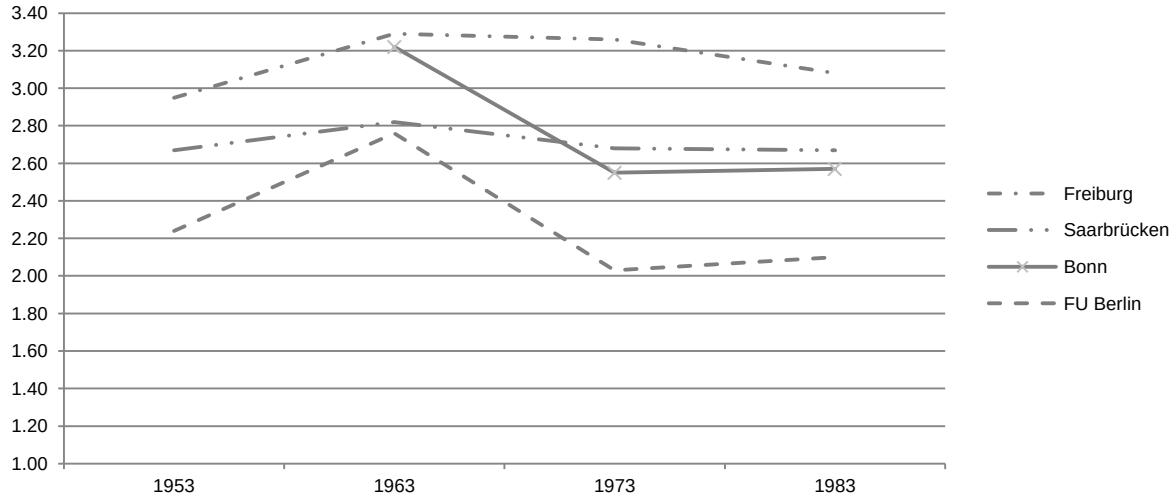
In Chemie zeigt sich ein Abwärtstrend vor allem an der Universität Würzburg, an der die Noten 1983 im Durchschnitt um 0.85 Noten besser ausfallen als noch 1953, und in geringerem Maße in Göttingen. In Gießen dagegen sinkt das Notenniveau zwischen 1953 und 1973 nur leicht und steigt dann 1983 bis über das ursprüngliche Niveau hinaus an. In Mathematik sinkt der Notendurchschnitt nur in Freiburg über alle drei Messzeitpunkte hinweg. Aachen, Berlin, Gießen und Göttingen weisen 1973 eine Verbesserung des Niveaus gegenüber 1963 und eine anschließende Verschlechterung 1983 auf, wobei sich in Gießen und Göttingen über den gesamten Zeitraum dennoch eine deutliche Verbesserung des Niveaus ergibt.

Abbildung 24: Abschlussnoten im Diplomstudiengang Mathematik an einzelnen Hochschulen – vier Zeitpunkte



Quelle: Hitpass/Trosien 1987, eigene Darstellung

Abbildung 25: Abschlussnoten im Diplomstudiengang VWL an einzelnen Hochschulen – vier Zeitpunkte



Quelle: Hitpass/Trosien 1987, eigene Darstellung

In VWL zeigt sich ein Anstieg für alle Universitäten bis 1963 (keine Daten für Bonn), gefolgt von einem Absinken bis 1973. 1983 ist dann in allen Fällen nur noch eine geringfügige Veränderung des Niveaus gegenüber 1973 festzustellen. Über den gesamten Zeitraum verbessert sich der Durchschnitt nur in Bonn deutlich, allerdings fehlen dort die Werte für 1953. In Berlin verbessert er sich leicht, während er in Freiburg etwas schlechter wird und in Saarbrücken wieder den Ausgangswert erreicht.

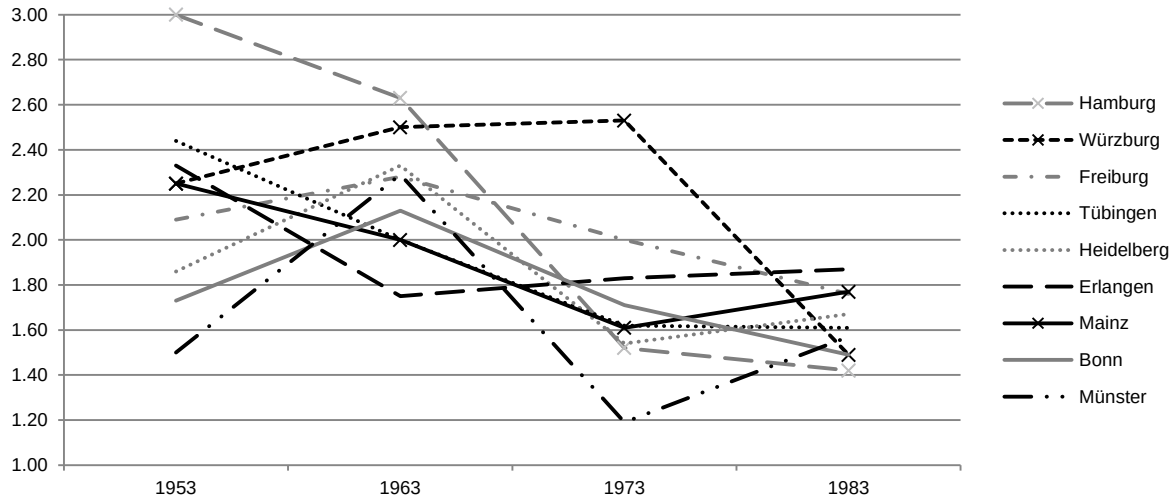
Für das Fach BWL enthält die Stichprobe von Hitpass und Trosien nur zwei Hochschulen, Berlin und Saarbrücken. Während der Notendurchschnitt in Saarbrücken über die vier Erhebungszeitpunkte hinweg sinkt, wird er in Berlin 1963 zunächst schlechter, 1973 besser und 1983 wieder schlechter, erreicht dort einen höheren Wert als 1953.

Abbildung 26: Abschlussnoten im Diplomstudiengang BWL an einzelnen Hochschulen – vier Zeitpunkte

Quelle: Hitpass/Trosien 1987, eigene Darstellung

Die neun in der Stichprobe enthaltenen Hochschulen mit Noten von Psychologieabsolvent*innen weisen schließlich sowohl unterschiedliche Entwicklungen, als auch einen gemeinsamen Trend auf: In Hamburg werden die Noten im Vergleich zu den früheren Messungen zu jedem Erhebungszeitpunkt besser, in Tübingen und Mainz ist dies bis 1973 der Fall, 1983 weisen die Noten in Tübingen denselben Durchschnitt auf wie 1973, in Mainz werden sie leicht schlechter. In Bonn und Freiburg sinken die Noten ab 1963, nachdem sie zuvor gegenüber 1953 angestiegen sind. In Heidelberg und Münster werden die Noten 1963 zunächst schlechter, verbessern sich 1973 und werden dann 1983 wieder schlechter. Erlangen und Würzburg stellen die deutlichsten Abweichungen im Verlauf dar. Während die Noten in Erlangen lediglich 1963 im Vergleich zu 1953 besser sind und nach 1963 wieder leicht schlechter werden, werden sie in Würzburg bis 1973 schlechter, fallen dann zu 1983 aber um eine komplette Note ab. Die Universität Münster stellt dahingehend die Ausnahme dar, als dass sie die einzige enthaltene Hochschule in Psychologie ist, an der das Notenniveau 1983 nicht unter dem von 1953 liegt. In allen anderen Fällen ist trotz des unterschiedlichen Verlaufs ein gemeinsamer Trend zu besseren Noten festzustellen.

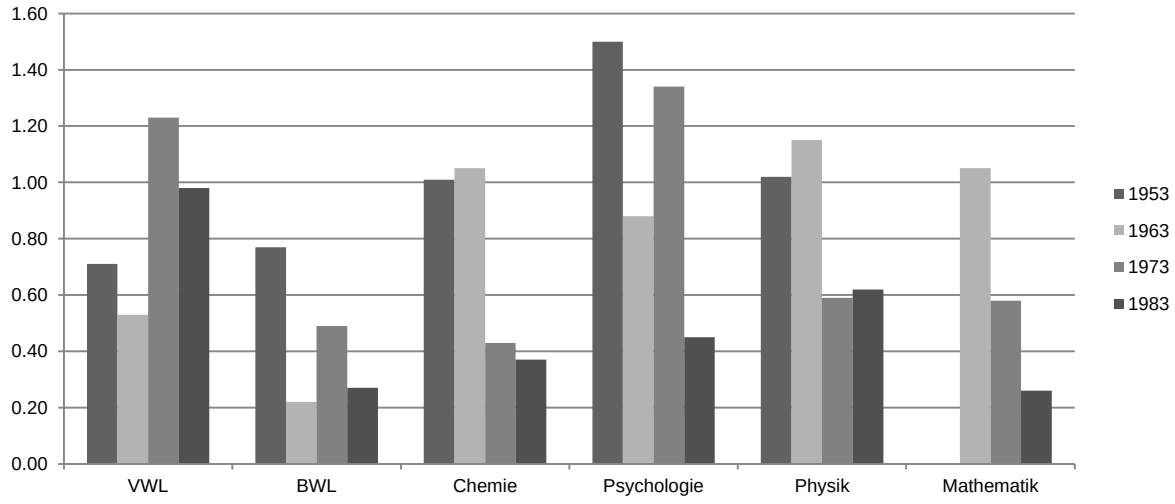
Abbildung 27: Abschlussnoten im Diplomstudiengang Psychologie an einzelnen Hochschulen – vier Zeitpunkte



Quelle: Hitpass/Trosien 1987, eigene Darstellung

Wird die Entwicklung der Unterschiede zwischen den Hochschulen innerhalb der einzelnen Studiengänge betrachtet, so fällt auf, dass sie 1983 überall, außer in VWL, geringer ausfallen als noch 1953. Eine kontinuierliche Angleichung im Zeitverlauf ist aus den Daten jedoch höchstens für Mathematik und eventuell für Chemie abzulesen.

Abbildung 28: Spannweiten zwischen den Hochschulen in sechs Diplomstudiengängen – vier Zeitpunkte

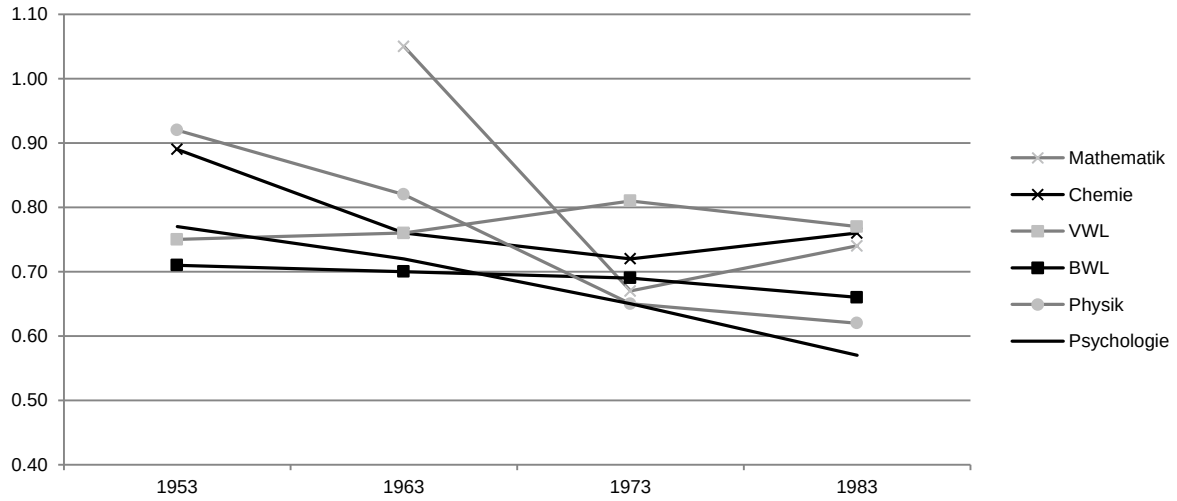


Quelle: Hitpass/Trosien 1987, eigene Darstellung

Hinsichtlich der Streuung der Noten zeigen die Daten von Hitpass/Trosien eine tendenzielle Verringerung im Zeitverlauf in fünf der sechs Diplomstudiengänge. In Psychologie, Physik und BWL sinkt die Standardabweichung über alle Messzeitpunkte hinweg. In Mathematik und Chemie sinkt sie bis 1973 und steigt 1983 wieder leicht an, lediglich in VWL liegt sie 1983 über dem Wert von 1953. Die Noten in den fünf Studiengängen mit Abwärtstrend scheinen sich also im Zeitverlauf näher an den Mittel-

wert zu bewegen, was bedeutet, dass das Notenspektrum dort zunehmend weniger ausgeschöpft wird[48].

Abbildung 29: Gemeinsame Standardabweichungen der Notenmittel in sechs Diplomstudiengängen – vier Zeitpunkte

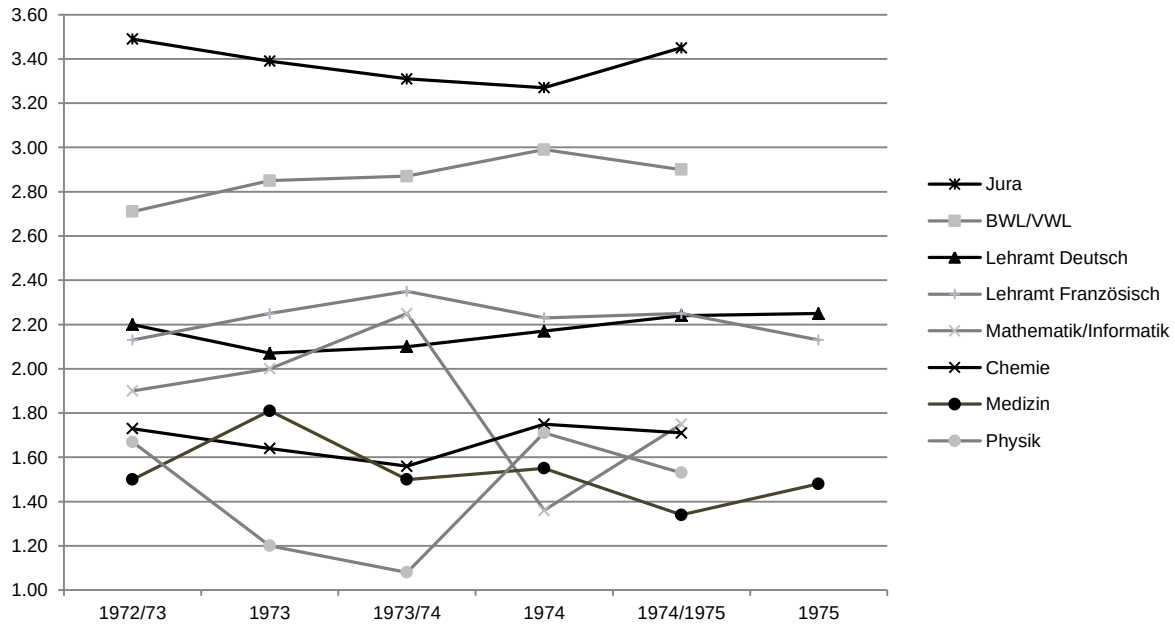


Quelle: Hitpass/Trosien 1987, eigene Berechnungen

Auch Apenburg et al. (1976) versuchen die an der Universität des Saarlandes erhobenen Noten im Zeitverlauf auszuwerten. Sie verfügen dazu im Gegensatz zu Hitpass/Trosien über einen wesentlich kürzeren Zeitabschnitt, aber über enger gefasste Messzeitpunkte. Auch wenn der Zeitraum, den Bemühungen der Autoren, in BWL/VWL „eine schwache Tendenz zu schlechteren Noten" (Apenburg et al. 1976:10), in Medizin „eine schwache Tendenz zur Verbesserung" zu erkennen, zum Trotz, keinesfalls ausreicht, um einen tatsächlichem Trend in der Notengebung in einem der Studiengänge zu identifizieren, bieten die Daten eine wertvolle Erkenntnis: Bereits innerhalb eines Semesterübergangs sind deutliche Schwankungen im Notenniveau zu beobachten, die im Falle von Mathematik/Informatik (Wintersemester 1973/74 zu Sommersemester 1974) fast bis zu einer ganzen Note (0.89) reichen. Dieser Umstand belegt, dass Vergleiche von Notenniveaus zu einzelnen gegebenen Zeitpunkten hinsichtlich der Ermittlung der Höhe fach-, hochschul- oder auch abschlussspezifischer Notenniveaus ebenso in die Irre führen können, wie dies bei einzelnen Messzeitpunkten in größeren Zeitabständen bezüglich der langfristigen Entwicklung von Notenniveaus der Fall sein kann.

[48] Auf eine Darstellung der Entwicklung der Streuung der Notenmittelwerte an den einzelnen Hochschulen wird wie auch bereits im Querschnittsvergleich aufgrund der teils sehr geringen Fallzahlen verzichtet.

Abbildung 30:Verlauf der Abschlussnoten an der Universität des Saarlandes vom WiSe 1972/73 bis max. SoSe 1975



Quelle: Apenburg et al. 1976, eigene Darstellung

Als der Wissenschaftsrat 2003 das Thema Hochschulnoten wieder aufgriff, war es vor allem die Presse, die, analog zu ihrem US-amerikanischen Pendant, einen Inflationsprozess in der Notenvergabe ausmachte. Interessanterweise wurde die Feststellung des Wissenschaftsrats (2003), dass 1996, 1998 und 2000 in den meisten Fächern vorwiegend gute Noten vergeben wurden und entsprechend gute Notendurchschnitte vorliegen in den Medien ohne jeden Zweifel als Folge eines inflationären Verbesserungsprozesses aufgefasst, obwohl zu dieser Einschätzung gar keine Vergleichszahlen zu früheren Zeitpunkten herangezogen wurden und im deskriptiven Arbeitsbericht auch keine Ursachen für die Vergabe guter Noten ermittelt wurden : „Viele Hochschulen vergeben fast nur noch Einsen und Zweien. Jetzt kritisiert der Wissenschaftsrat die Inflation der Spitzenzensuren“ (Spiewak 2003; siehe auch Elias 2003).

Auch die 2007 erfolgte Feststellung des Wissenschaftsrats, dass an deutschen Hochschulen ein „über mehrere Prüfungsjahre (…) vergleichsweise einheitliches Notenniveau, häufig verbunden mit überwiegend guten Noten“ (Wissenschaftsrat 2007:11) festzustellen sei, wurde in der medialen Wahrnehmung auf diese Weise umgedeutet. Gerne zitiert wurde dabei ein im Vorwort verwendeter Satzteil, in dem ein „inflationärer Anstieg des Notenniveaus“ (ebd.) beschrieben wird (etwa bei Friedmann/Hinrichs 2007). Dass diese Passage jedoch nur auf nicht weiter spezifizierte „Beobachtungen“ (Wissenschaftsrat 2007:11) einer solchen Entwicklung als Hintergrund der Diskussion um Hochschulnoten verweist und nicht auf die tatsächlich vorhandene Datenlage, stört die Verfasser*innen dieser Artikel nicht weiter.

In seinem jüngsten Bericht zur Notenvergabe an deutschen Hochschulen aus dem Jahr 2012 lieferte der Wissenschaftsrat dann die Zeilen, die eine Verbesserung des Notenniveaus implizieren, an der in den Medien auch zuvor schon keine Zweifel bestanden:

> „Ein weiteres zentrales Ergebnis des vorliegenden Arbeitsberichtes ist die fortgesetzte Tendenz zur Vergabe besserer Noten. In den universitären Studiengängen mit traditionellen Abschlüssen – Diplom und Magister sowie Staatsexamen ohne Lehramt – ist beispielsweise der Anteil der mit „gut" oder „sehr gut" bewerteten Abschlussprüfungen zwischen 2000 und 2011 um knapp neun Prozentpunkte von 67,8 % auf 76,7 % gestiegen" (Wissenschaftsrat 2012:7).

Das mediale Echo auf diesen Befund war wie zu erwarten eindeutig: Aus der „fortgesetzten Tendenz zur Vergabe besserer Noten" (ebd.) wurde die „Tendenz zu immer besseren Prüfungsnoten" (Friedmann 2012), die Rede war von „Kuschelnoten" (Schlicht 2012) und natürlich von: Noteninflation, und zwar von „schleichender Noteninflation" einhergehend mit einer „Aufweichung der Bewertungsstandards" (Preuss 2012) – zitiert einmal mehr aus dem Vorwort des Arbeitsberichts, in dem der Vergleich der Prüfungsjahre 2000 und 2012 angekündigt wird. Im weiteren Bericht, auch im Abschnitt, in dem der angekündigte Vergleich der Prüfungsnoten vorgenommen wird, lassen sich derartige Formulierungen oder sonstige Hinweise auf eine Noteninflation nicht mehr finden. Verstärkt wurde der öffentlich Eindruck der allgemeinen Entwertung von Noten von Kommentaren des Wissenschaftsrats selbst (ebd.)[49].

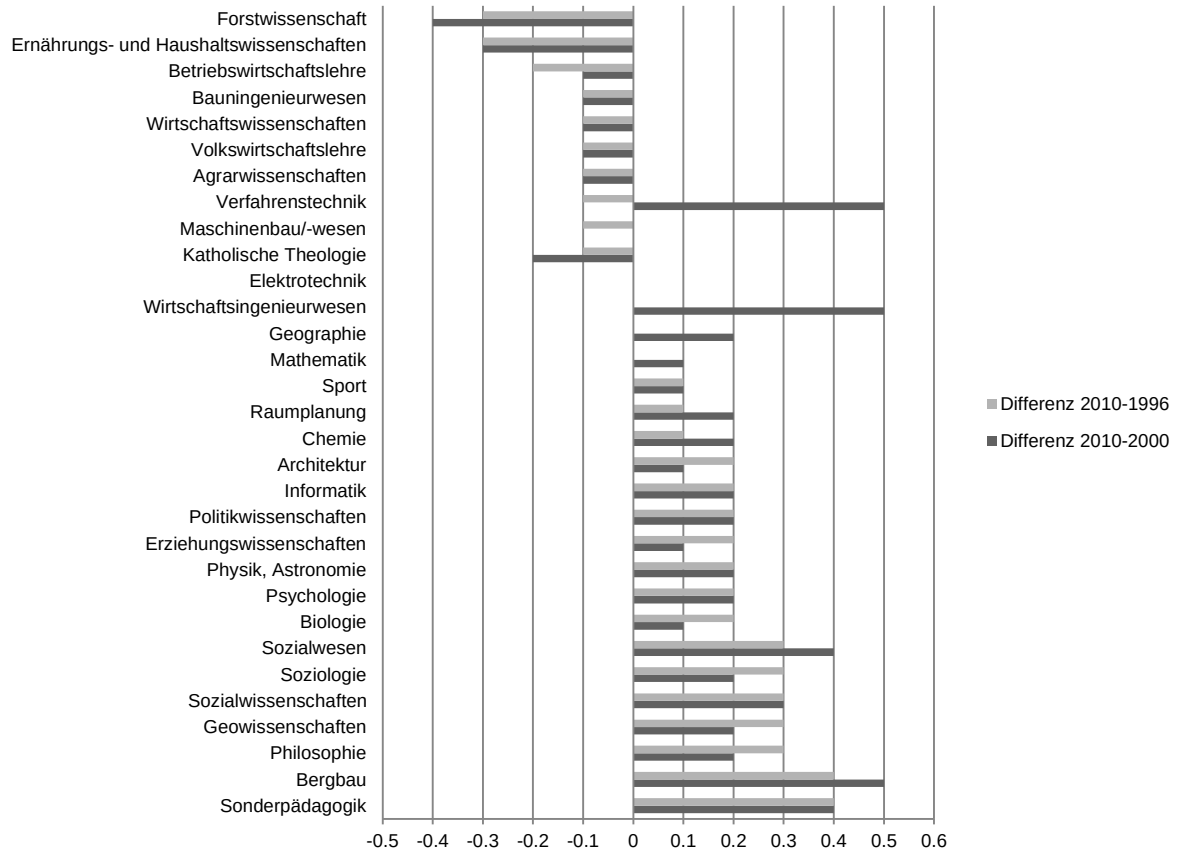
Angemerkt werden sollte zu vorigem Zitat und den darauf folgenden Reaktionen jedoch, dass der Wissenschaftsrat selbst 2007 noch darauf hinweist, dass die für das Prüfungsjahr 2005 „vorliegenden Ergebnisse (…) nicht mit dem im Jahr 2003 vorgelegten Arbeitsbericht für das Prüfungsjahr 2000 vergleichbar" (Wissenschaftsrat 2007:9) sind, in dem die Datenbasis als lückenhaft beschrieben wird und zu dem methodische Differenzen in den nachfolgenden Berichten bestehen. Wird dann noch die Mühe aufgebracht, mehr als die Zusammenfassung des Arbeitsberichtes zu lesen, wird wiederum schnell deutlich, dass die Diagnose Noteninflation möglicherweise verfrüht getroffen wurde. So klingt das Ergebnis der nach Fächergruppen differenzierten Analyse der Notenentwicklung (alle Abschlüsse außer Promotionen umfassend) kaum noch so spektakulär, wie die Feststellung des deutlichen Anstiegs guter und sehr guter Noten:

> „Im Vergleich der Notendurchschnitte der Prüfungsjahre 2000 und 2011 zeigt sich, dass sich die Noten in den Fächergruppen Humanmedizin/Gesundheitswissenschaften, Rechts-, Wirtschafts- und Sozialwissenschaften sowie Sprach- und Kulturwissenschaften von 2000 auf 2011 geringfügig verbessert haben. In den übrigen Fächergruppen sind die Notendurchschnitte etwa gleich geblieben bzw. haben sich minimal verschlechtert." (Wissenschaftsrat 2012:46)

[49] Dass das mediale Interesse an einer inflationären Notenentwicklung wesentlich größer ist als die zum Thema verfügbaren, belastbaren Erkenntnisse ist kein spezifisch deutsches Phänomen wie eine kurze Recherche von Oleinik (2009) zeigt: Seit 1993 findet Oleinik 84 wissenschaftliche Publikationen zum Thema grade inflation, nur 56 davon beinhalten empirische Ergebnisse und/oder theoretische Erklärungsansätze. Im Vergleich dazu ermittelt er für denselben Zeitraum 989 Presseartikel, die sich mit dem Thema befassen.

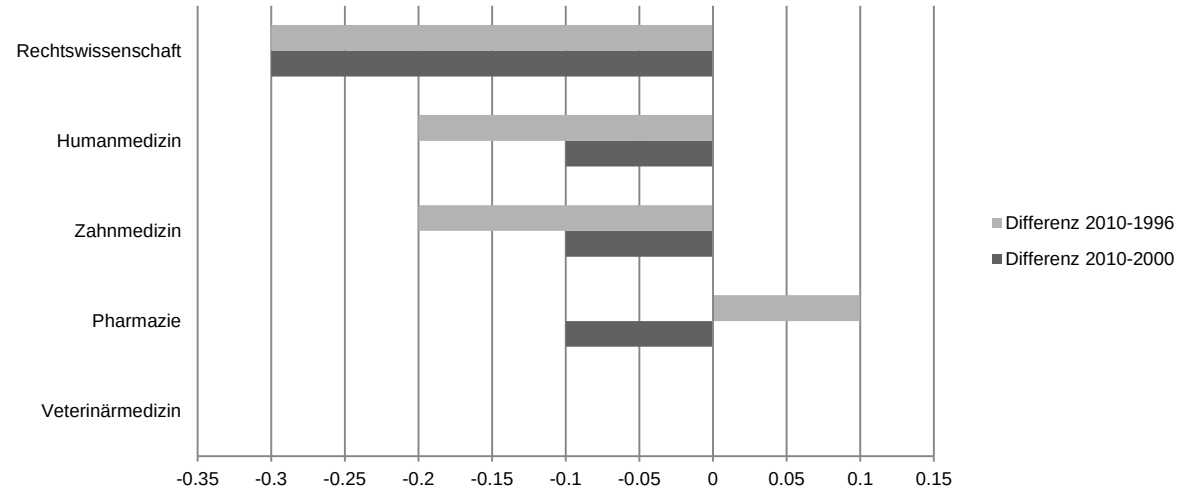
Werden nun die Veränderungen in den einzelnen Fächern in den Diplom- und Magisterabschlüssen betrachtet, zeigt sich ein Bild, das dem einer umfassenden Noteninflation im Zeitraum zwischen 1996 bzw. 2000 und 2010 deutlich entgegensteht: Lediglich in 10 von 31 Diplomstudiengängen an Universitäten, für die in allen drei Arbeitsberichten Zahlen vorliegen, hat sich das Notenniveau zwischen 1996 und 2010 verbessert (acht von 31 zwischen 2000 und 2010), und das in sieben (fünf) Fällen um gerade einmal 0.1 Noten. An Fachhochschulen kann für diesen Zeitraum nur in zwei von 19 Diplomstudiengängen (einer von 19 zwischen 2000 und 2010) eine Verbesserung festgestellt werden. Studiengänge mit Abschluss Magister weisen in keinem einzigen von 13 Fällen eine Verbesserung zwischen 1996 und 2010 auf (einer von 13 zwischen 2000 und 2010), solche mit Abschluss Staatsexamen immerhin in drei von fünf Fällen (vier von fünf zwischen 2000 und 2010), wobei zu beachten ist, dass in den Staatsexamensprüfungen durchschnittlich wesentlich schlechtere Ergebnisse erzielt werden als in den übrigen Abschlussarten.

Abbildung 31: Veränderungen im Notenniveau in ausgewählten Diplomstudiengängen (nur Universitäten)



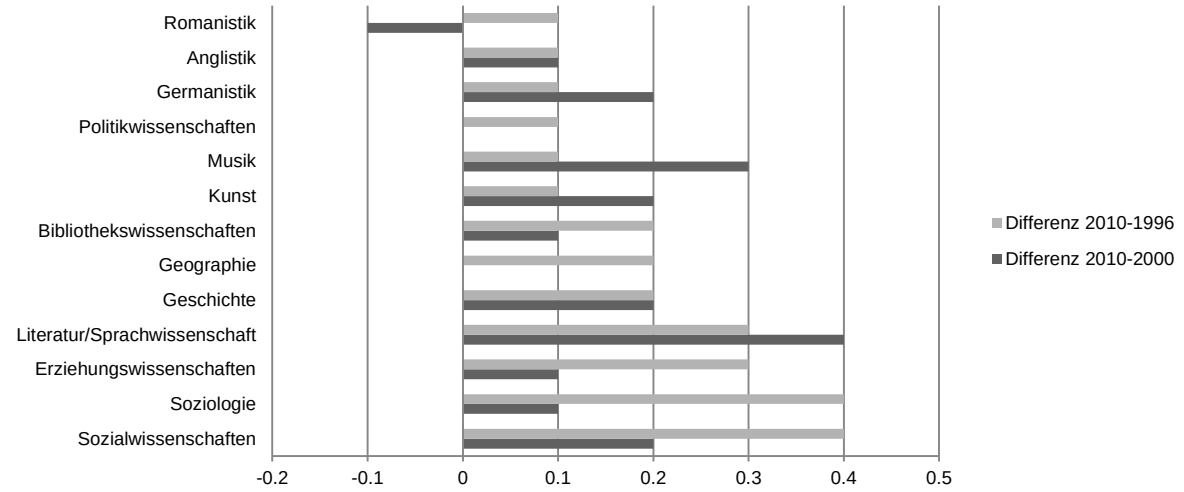
Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Berechnungen

Abbildung 32: Veränderungen im Notenniveau in ausgewählten Staatsexamensstudiengängen



Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Berechnungen

Abbildung 33: Veränderungen im Notenniveau in ausgewählten Magisterstudiengängen

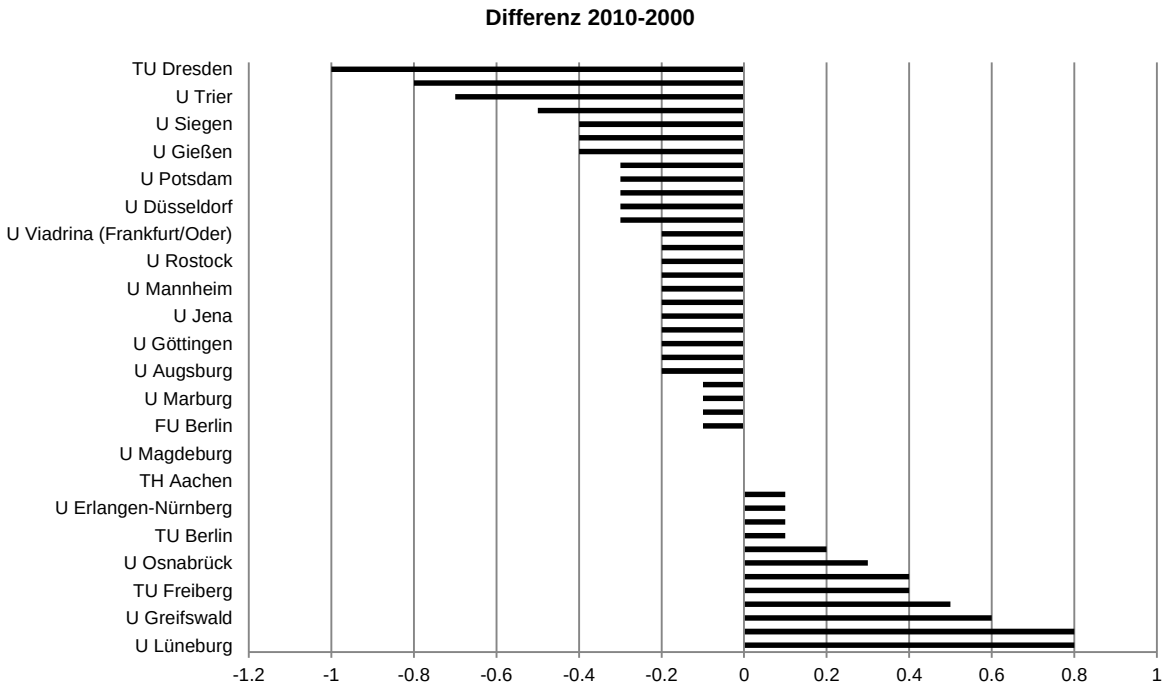


Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Berechnungen

Dass die Studiengangebene dabei nicht als alleinige Analyseebene für die langfristige Entwicklung von Noten betrachtet werden sollte - zu groß sind die Differenzen zwischen einzelnen Hochschulen innerhalb eines Studiengangs, die auf diese Weise vernachlässigt würden - lässt sich exemplarisch am Beispiel Betriebswirtschaftslehre illustrieren: Während die Noten auf Fachebene 2010 im Durchschnitt um 0.1 Noten besser liegen als 2000, verbesserten sie sich in diesem Zeitraum an der TU Dresden um eine ganze Note. An den Universitäten Greifswald und Lüneburg hingegen fallen die Abschlussnoten 2010 im Durchschnitt um 0.8 Noten schlechter aus als 2000[50].

[50] Die Daten von Krempkow, die zwar einen kürzeren Zeitraum abbilden als die Wissenschaftsratsberichte, dafür aber zwei Datenpunkte mehr und zudem geringere Abstände zwischen den Messungen aufweisen, weisen ebenfalls auf eine hochschulspezifische Vielfalt der Notenentwicklung für viele Fächer an den betrachteten sächsischen Universitäten hin (Krempkow 2002-2005).

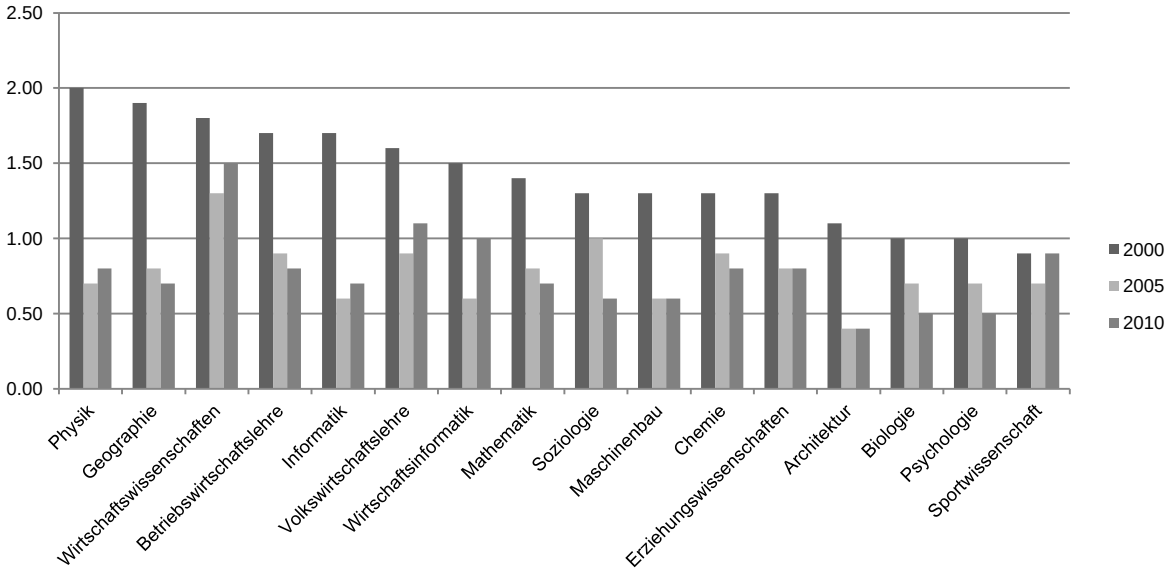
Abbildung 34: Veränderungen im Notenniveau im Diplomstudiengang Betriebswirtschaftslehre



Quelle: Wissenschaftsrat 2003; 2012, eigene Darstellung

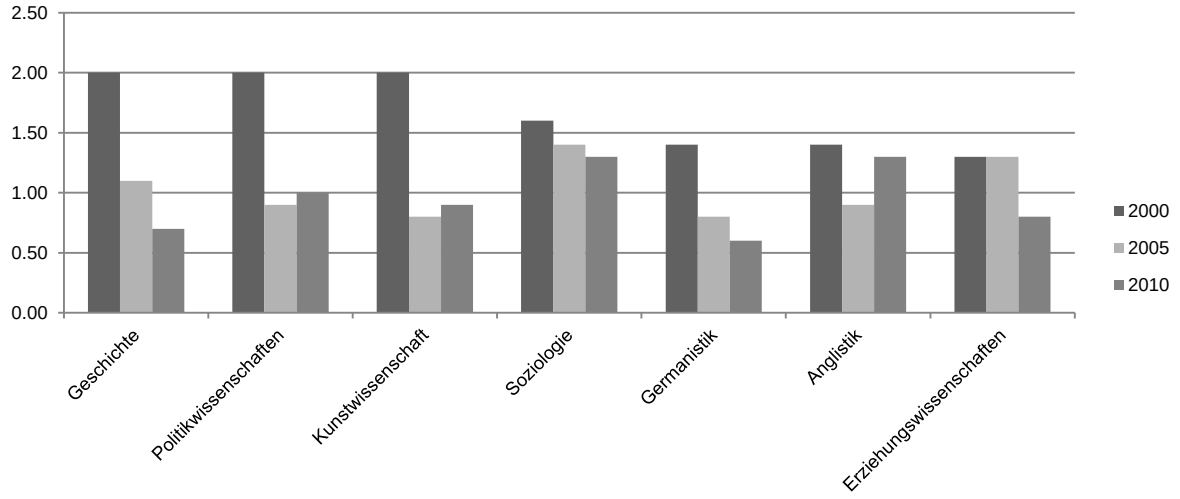
Auch die Spannweiten zwischen den Hochschulen innerhalb der gleichen Studiengänge (Abb.35 und 36) sinken im Zeitraum der Berichterstattung durch den Wissenschaftsrat. Es fällt allerdings bei Betrachtung der Hochschulspannweiten auf, dass sich die größte Annäherung der Extremwerte zwischen 2000 und 2005 vollzieht, während zwischen 2005 und 2010 auch leichte Vergrößerungen zu beobachten sind. Dies ist möglicherweise auf methodische Differenzen in der Erhebung der Noten zwischen den einzelnen Berichten zurückzuführen.

Abbildung 35: Spannweiten zwischen den Hochschulen in ausgewählten Diplomstudiengängen – drei Zeitpunkte



Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Berechnungen

Abbildung 36: Spannweiten zwischen den Hochschulen in ausgewählten Magisterstudiengängen - drei Zeitpunkte



Quelle: Wissenschaftsrat 2003; 2007; 2012, eigene Berechnungen

Die Differenz zwischen den am besten und am schlechtesten benotenden Fächern (über alle „traditionellen" Abschlüsse, also ohne Bachelor und Master, und alle Hochschulformen hinweg) sinkt von 2.0 Noten in 1996 (2.2 in 1998; 2.1 in 2000) auf 1.8 Noten in 2005 bzw. 1.5 Noten in 2010[51], während die Abschlussunterschiede konstant bleiben: In sechs der 36 universitären Diplom- und 16 Magisterstudiengänge beläuft sich die Durchschnittsnote 2005 auf schlechter als $\bar{x}$=2.0 - während dies in drei der vier aufgeführten Staatsexamensstudiengänge der Fall ist (ebd. 2007). 2010 ist die Durchschnittsnote in sechs von 23 Diplom- und acht berücksichtigten Magisterstudiengängen schlechter als $\bar{x}$=2.0, was auf drei der fünf aufgeführten Staatsexamensstudiengänge zutrifft (ebd. 2012).

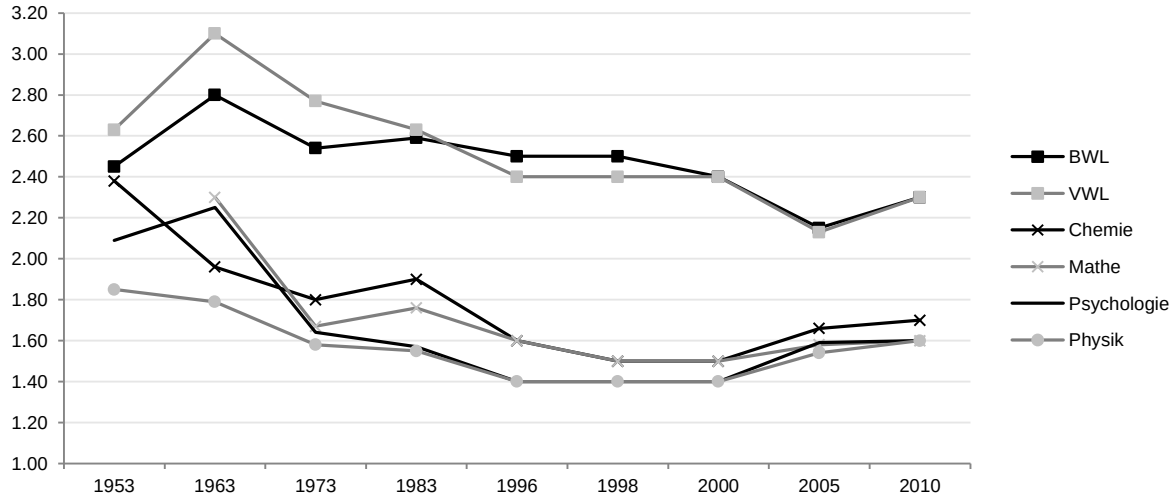
Im Zeitraum zwischen 1996 und 2010 ist eine Notenverbesserung an deutschen Hochschulen auf Fachebene also nur ein vereinzelt auftretendes Phänomen. Doch wie sieht es mit dem Zeitraum vor 1996 aus? Bauer und Grave (2011) präsentieren Notendurchschnitte aus dem Konstanzer Studierendensurvey für die im Survey enthaltenen Hochschulen aus Nordrhein-Westfalen, Baden-Württemberg, Bayern, Hessen und Sachsen, aggregiert auf Bundeslandebene.

Sie stellen für alle Bundesländer außer Sachsen eine Verbesserung der Noten von 1983 bis 2007 fest. Die Autor*innen geben dabei an „the average grade of student *i* at time *t* they have earned during their undergraduate study" (Bauer/Grave 2011:7, Herv. i.O.) zu nutzen - der Konstanzer Studierendensurvey enthält laut Codebuch allerdings nur die Abitur- und die Zwischenprüfungsnote sowie alternativ eine Selbsteinschätzung der bisherigen Studienleistungen in Notenform, falls zum Zeitpunkt der Befragung noch keine Zwischenprüfung abgelegt wurde. Diese Daten als aussagekräftige Durchschnittsnoten zu verwenden, ist zumindest diskussionswürdig. Mehr Aufschluss über die lang-

[51] Für 2000 wurde der Studiengang Bildende Kunst mit einem Wert von 0.8 nicht berücksichtigt. Die Spannweite bezieht sich nur auf die Fächer, die in allen drei Berichten enthalten sind, um Vergleichbarkeit zu gewährleisten.

fristige Notenentwicklung an deutschen Hochschulen bietet ein Vergleich der Noten aus den Arbeitsberichten des Wissenschaftsrates mit denen von Hitpass und Trosien (1987). Hier zeigt sich für die sechs vergleichbaren Studiengänge tatsächlich eine Verbesserung der Noten vom ersten Messzeitpunkt 1953 bis zum letzten Messzeitpunkt 2010. Nach 1963 sinkt das Notenniveau in allen Studiengängen tendenziell, bevor es sich ab 1996 stabilisiert und dann (in BWL und VWL erst ab 2005) wieder leicht ansteigt.

Abbildung 37: Abschlussnoten in sechs ausgewählten Fächern - neun Zeitpunkte



Quelle: Hitpass/Trosien 1987; Wissenschaftsrat 2003;2007;2012, eigene Darstellung

Eine langfristige Verbesserung der Notenniveaus könnte demnach, zumindest in einzelnen Studiengängen, tatsächlich stattgefunden haben und zwar bereits lange bevor die öffentliche Diskussion darüber in Tritt kam[52].

Doch wie könnte eine langfristige Notenverbesserung an deutschen Hochschulen erklärt werden? Einzelne Stimmen verweisen in Reaktion auf die vermeintliche Inflationswelle der letzten 20 Jahre auf bereits bekannte Argumente aus der US-amerikanischen Forschung. Der Tausch guter Evaluationen gegen gute Noten, ein freundschaftliches Verhältnis zwischen Lehrenden und Lernenden, übermäßige Arbeitsbelastung der Lehrenden, die Vermeidung von Beschwerden, sowie die Weigerung, Absolvent*innen (auch wegen schlechter Lehrbedingungen) mit schlechten Abschlüssen auf den Arbeitsmarkt zu schicken werden intuitiv als (von ihren Verfasser*innen nicht geprüfte) Einflussfaktoren genannt (Bayer 2013 fächergruppenspezifisch für die Geisteswissenschaften; Elias 2003; Kühl 2012; Unispiegel 2007).

An anderer Stelle lassen sich auch vereinzelte empirische Begründungen finden, die etwa darauf hinweisen, dass zunehmende Zweifel an der Aussagekraft von Noten zu einer nachsichtigeren Benotung geführt haben könnten. So berichten Hochschullehrende, nach Veränderungen in der Bewer-

[52] Die Arbeit von Müller-Benedict et al. (2008) deutet darauf hin, dass (in den klassischen akademischen Karrieren) bis 1941 noch keine langfristigen Notenverbesserungen zu verzeichnen waren. Ein Beginn derartiger Entwicklungen ist daher ab Gründung der Bundesrepublik möglich.

tungspraxis befragt, von Auswirkungen der Studierendenbewegung der 1960er Jahre auf die Wahrnehmung von Prüfungen und Notengebungsprozessen, etwa hinsichtlich der „Fragwürdigkeit der herkömmlichen Leistungsmessung“ (Ziegenspeck 1999:259).
Müller-Benedict und Tsarouha (2011) verweisen, wie in Abschnitt 6.2 bereits ausführlich erläutert, auf den möglichen Einfluss des Akademiker*innenarbeitsmarktes auf die Höhe des Notenniveaus. Demnach wäre es denkbar, dass sich das Notenniveau mit zunehmend schlechterer Arbeitsmarktlage ebenfalls verschlechtert, weil sich Prüfende in Zeiten der akademischen Überfüllung eher wieder der Selektionsfunktion des Bildungssystems bewusst werden. Die Autor*innen vermuten, ausgehend vom Befund zyklischer Mangel- und Überfüllungsperioden im Akademiker*innenarbeitsmarkt (Titze 1990) zyklische Auswirkungen auf die Höhe des Notenniveaus (siehe auch Müller-Benedict 2005 zum Zusammenhang zwischen Akademiker*innenzyklen und Prüfungserfolg). Im deutschen Kontext sind Müller-Benedict und Tsarouha die ersten, die explizit auf die Möglichkeit zyklischer Notenverläufe hinweisen. Über kurze Zeiträume beobachtete Verbesserungsperioden stellen dieser Überlegung nach nicht zwangsläufig einen kontinuierlichen Trend zu besseren Noten dar, sondern sind möglicherweise nur Bestandteil einer umfassenderen Dynamik. In den USA wird die Möglichkeit einer zyklischen Prozessdynamik nur von Kolevzon (1981) in Betracht gezogen.

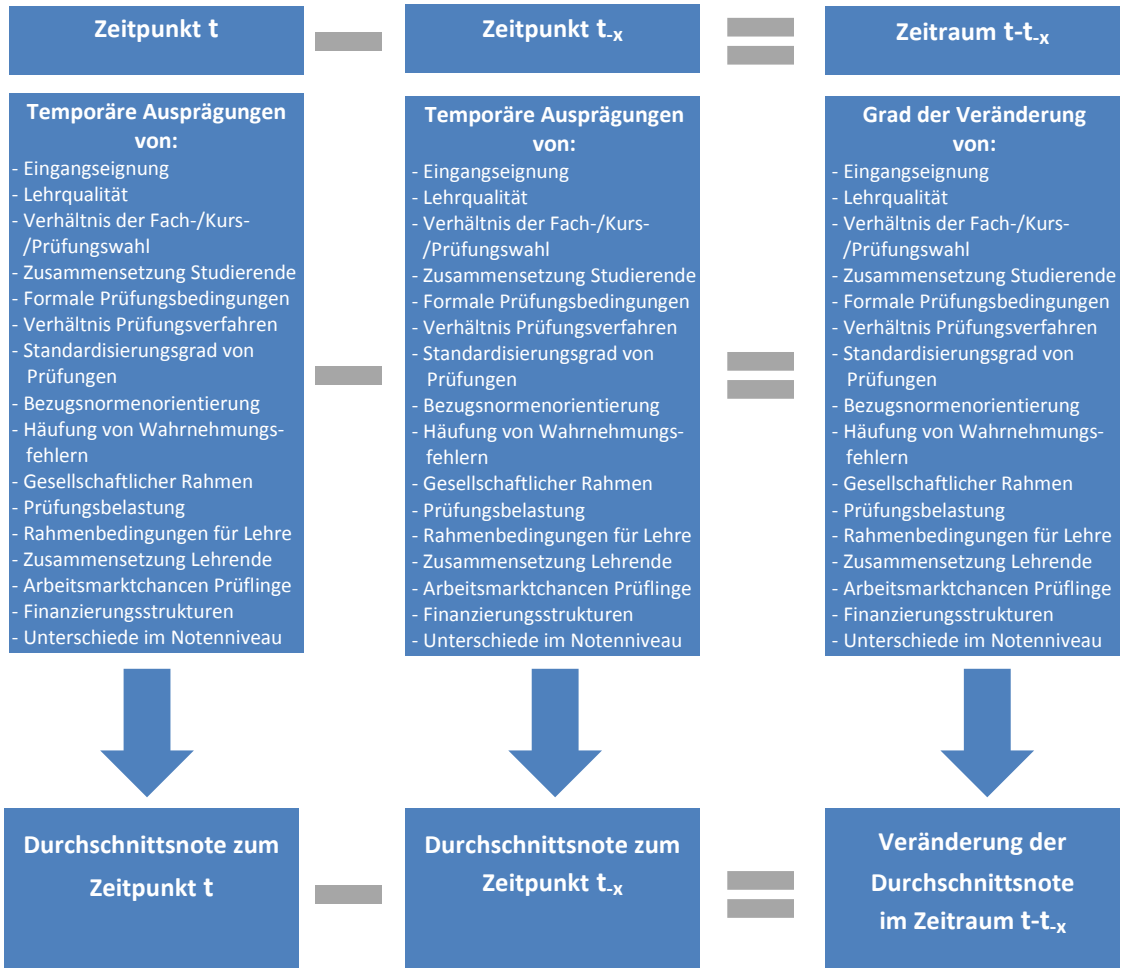
Die Betrachtung der verfügbaren Daten zur Notenentwicklung an deutschen Hochschulen zeigt, dass die Noten sich in den letzten Jahren nur in wenigen Studiengängen zum Besseren entwickelt haben und die wahrgenommene Noteninflation die tatsächliche Verbesserung bei Weitem übersteigt. Sie lassen erkennen, dass auch die Entwicklungspfade des Notenniveaus sowohl fach- als auch hochschul- und möglicherweise abschlussspezifisch analysiert werden müssen. Die Daten von Hitpass und Trosien weisen darauf hin, dass die Streuung der Noten im Zeitverlauf abnimmt, können jedoch aufgrund der großen Abstände zwischen den Messzeitpunkten nur bedingt als Beleg für eine abnehmende Ausnutzung der Breite der Notenskala dienen.
Durch die Verknüpfung der Daten des Wissenschaftsrates mit denen von Hitpass und Trosien deutet sich in sechs betrachteten Studiengängen eine langfristige Verbesserung des Notenniveaus seit 1963 an. Allerdings sind die Messzeitpunkte, für die die Daten auf diese Weise zur Verfügung stehen, zu wenige, um von ihnen auf eine kontinuierliche Entwicklung schließend zu können. Auch wenn mehrere nicht unmittelbar aufeinanderfolgende Zeitpunkte einzeln oder aufeinanderfolgende Zeitpunkte gemittelt erfasst werden, besteht die Möglichkeit, das bestimmte Prozessdynamiken nicht erfasst werden, die im einen Fall zwischen den Zeitpunkten „lauern“, im anderen Fall durch Aggregation verdeckt werden. Deshalb sind, wie bereits erläutert, lange, durchgehende Zeitreihen nötig, die garantieren, dass in keinem Studiengang und an keiner Hochschule zu irgendeinem der betrachteten Zeitpunkte ein Ausreißerwert als Hinweis auf einen Trend interpretiert wird.

Sollten sich an deutschen Hochschulen langfristige Verbesserungen des Notenniveaus feststellen lassen, kommen, unter Berücksichtigung nationaler Besonderheiten des Hochschulsystems, die gleichen Ursachen für dieses Phänomen in Frage, die auch in der internationalen Literatur diskutiert werden. Tabelle 6 ordnet alle in der Literatur besprochenen potentiellen Gründe für Notenverbesserungen im Zeitverlauf den zuvor erstellten Ursachenkategorien zu. Abbildung 38 gibt einen grafischen Eindruck des modellhaften Verlaufs der Notenentwicklung, nach dem die Differenz zwischen den durchschnittlichen Prüfungsnoten zweier Messzeitpunkte auf die Differenz zwischen den zu den jeweiligen Messzeitpunkten herrschenden Leistungspotentialen und Prüfungsbedingungen zurückgeführt werden kann.

Tabelle 6: Einordnung der besprochenen Ursachen für gesunkene Bewertungsstandards in Ursachenkategorien

In der Literatur angeführte Ursachen für gesunkene Bewertungsstandards	Übergeordnete Ursachenkategorien
- gestiegene Wahlmöglichkeiten - Freiversuche, Bestehen ohne Note, vereinfachter Rücktritt	- Änderungen formaler Prüfungsbedingungen
- gestiegener Anteil schriftlicher Leistungen (vs. mündliche)	- Änderungen in den eingesetzten Prüfungsverfahren
- gestiegener Anteil standardisierter Tests (vs. Aufsätze und mündliche Prüfungen)	- zunehmende Standardisierung von Prüfungen
- Wechsel von sozialer zu sachlicher Bezugsnorm	- veränderte Bezugsnormenorientierung
- Absicherung gegen Wehrdienst - Fördermittel für Minderheiten - Wirtschaftliche Entwicklung - Zweifel an der Aussagekraft von Noten	- gesellschaftliche Ereignisse und Entwicklungen
- Mittel, um gestiegene Prüfungsbelastung zu verringern	- gestiegene Prüfungsbelastung
- „disengagement compact“ (Kuh 2003:28) - Ausgleich für verschlechterte Betreuungsrelationen - studentische Evaluation der Lehrleistung	- veränderte Rahmenbedingungen für Lehre
- Arbeitsverhältnis, politische und pädagogische Einstellungen, Alter, Generation, Geschlecht, Ethnizität, Forschungserfolg	- veränderte Zusammensetzung der Lehrenden
- Mangel auf dem Arbeitsmarkt - zunehmende Polarisierung zwischen Arbeitsplätzen mit hohem und niedrigem Anforderungsprofil - zunehmender Konkurrenzkampf um wenige gute Stellen bei immer mehr geeigneten Absolvent*innen	- veränderte Arbeitsmarktchancen der Prüflinge
- Noten als Steuerungsinstrument, um Förderung/Reputation zu erhöhen - Noten als Gut für zahlende Kunden (gestiegene Studiengebühren)	- veränderte Finanzierungsstrukturen
- Anpassung an bessere Vergleichsniveaus, um eigene Studierende nicht zu benachteiligen	- unterschiedliche Notenniveaus

Abbildung 38: Veränderung von Leistung und Prüfungsbedingungen als Ursache für Veränderungen im Notenniveau



7. Datenbasis

Da die Prüfungsstatistik deutscher Hochschulen erst seit dem Wintersemester 1992/93 geführt wird (und dies erst ab ca. 1995 weitestgehend vollständig), stehen amtliche Daten, die die Examensnoten von Studierenden beinhalten nur für einen relativ kurzen Zeitraum zur Verfügung. Die einzige Studie, die sich umfassend mit der Notengebung an deutschen Hochschulen vor 1995 befasste (Hitpass/Trosien 1987) konzentrierte sich auf wenige Zeitpunkte, für die die Noten erhoben und miteinander verglichen wurden. Problematisch an diesem Vorgehen ist, dass nicht überprüft werden kann, ob einzelne Zeitpunkte Ausreißerwerte oder Extrempunkte von Zyklen darstellen. Ein Trend oder gar ein charakteristischer Verlauf ist aus solchen Daten streng genommen nicht abzulesen, kann höchstens erahnt werden.

Die im Folgenden verwendeten Daten beheben diesen Mangel. Sie entstammen dem Forschungsprojekt „Die Notengebung an Hochschulen in Deutschland von den 1960er Jahren bis heute. Trends, Unterschiede, Ursachen.“. Im Rahmen dieses Projekts wurden die Abschlussnoten der Studierenden der Georg-August Universität Göttingen, der Technischen Universität Braunschweig, des Karlsruher Instituts für Technologie (ehemals Technische Hochschule Karlsruhe), der Freien Universität Berlin, der Eberhard Karls Universität Tübingen, der Ruprecht-Karls-Universität Heidelberg, der Westfälischen Wilhelms-Universität Münster und der Universität des Saarlandes (nur Germanistik) der letzten Jahrzehnte in ausgewählten Studiengängen erhoben. Der früheste Zeitpunkt, bis zu dem die Noten erhoben werden konnten, variiert dabei von Hochschule zu Hochschule, zum einen nach Möglichkeiten und Interesse, Prüfungsakten bzw. Zeugnisse über die gesetzlich festgelegte Aufbewahrungsfrist hinaus zu verwahren, zum anderen nach praktischer Verfügbarkeit.

Die Auswahl der in der Erhebung berücksichtigten Hochschulen und Studiengänge wurde zunächst theoretisch, dann forschungspragmatisch vorgenommen. Es wurden Hochschulen ausgewählt, die die Überprüfung regionaler (z.B. Bundesland politisch/ Stadtstaat vs. Flächenstaat) sowie hochschulspezifischer Unterschiede (z.B. Alter/Größe) zulassen und als Standorte für Gruppendiskussionen, die im qualitativen Projektbereich geführt wurden, geeignet sind. Die Auswahl der Studiengänge sollte das Fächergruppenspektrum abbilden und die Überprüfung von Unterschieden nach Zugangsvoraussetzungen, der Heterogenität der Studieninhalte (z.B. standardisierter Studienverlauf vs. große Auswahl) und nach Abschlussart (Staatsexamen vs. Diplom/Magister) ermöglichen.

In den Daten enthalten sind die Studiengänge Betriebswirtschaftslehre, Volkswirtschaftslehre, Psychologie, Mathematik, Biologie, Chemie, Maschinenbau (alle mit Abschluss Diplom), Soziologie (Magister und Diplom), Germanistik (Magister) sowie Mathematik und Deutsch als Lehramtsstudium (Abschluss 1. Staatsexamen), soweit sie an den ausgewählten Hochschulen angeboten wurden und die Daten dort verfügbar waren. Die Zuordnung der Prüflinge zu den Diplomstudiengängen stellte kein Problem dar, da die Abschlussprüfung dort klar einem Studiengang zugeordnet ist. In den Magis-

ter- und Lehramtsstudiengängen hingegen können die Studierenden oft zwei gleichwertig geprüfte Hauptfächer belegen. Hier erfolgte die Zuordnung zu den in der Stichprobe enthaltenen Fächern immer dann, wenn sie (Deutsch/Mathematik im Lehramt; Germanistik/Soziologie im Magister) als erstes *oder* zweites Hauptfach belegt wurden. Um Dubletten im Datensatz zu vermeiden, wurden Prüflinge, die in beiden Fächern gleichwertig geprüft wurden, nur dem Fach zugeordnet, in dem sie als erstes Hauptfach angemeldet waren. Magister- und Lehramtsabsolvent*innen, die nur ein Hauptfach belegt haben und dazu in zwei Nebenfächern geprüft wurden, wurden nur den im sample enthaltenen Studiengängen zugeordnet, wenn sie sie als Hauptfach gewählt hatten.

Tabelle 7 gibt einen Überblick, für welche Jahre Daten in den einzelnen Studiengängen an welchen Hochschulen vorliegen. Die Noten wurden an den Hochschulen bis 1997 aufgenommen und anschließend mit den Daten der Prüfungsstatistik verknüpft, aus der dazu die jährlichen Mittelwerte (Variable ef133: „Jahr des Prüfungsabschlusses) der Noten der im sample enthaltenen Hochschulen berechnet wurden[53]. Der Zugang zur Hochschulprüfungsstatistik erfolgte über das Forschungsdatenzentrum (FDZ) des Statistischen Landesamts Schleswig-Holstein (2012). Dadurch stehen auf Hochschulebene Zeitreihen der Abschlussnoten im Umfang von maximal 68, auf Studiengangebene von maximal 61 Jahren zur Verfügung (Die Noten wurden nur für die Jahre zu einem Durchschnitt auf Studiengangebene zusammengefasst, für die von mindestens zwei Hochschulen Werte vorliegen). Außerdem enthalten die Daten die Prüfungsergebnisse der Absolvent*innen der Rechtswissenschaften als aggregierte Durchschnittsnoten auf Bundeslandebene von 1959 bis 2007 (1. Staatsexamen) bzw. bis 2011 (2. Staatsexamen). Diese konnten aus der juristischen Fachzeitschrift „Juristische Schulung“ erhoben werden. Dort werden die Prüfungsergebnisse aller Prüfungsdurchgänge der Prüfungen zum ersten und zweiten Staatsexamen seit 1959 jährlich veröffentlicht.

Die Zahl der Prüflinge in den einzelnen Studiengängen reicht in den verwendeten Daten des Hochschulsamples von n=9 (Mathematik Lehramt/2010) bis n=1695 (BWL/1993) pro Jahr, an den einzelnen Hochschulen von n=1 (Germanistik/Tübingen/1972) bis n=618 (BWL/Münster/ 1993). In der bundesweiten Totalerfassung der ersten juristischen Staatsprüfungen spannt sich die Fallzahl von N=2698 (1965) bis N=11424 (1996) insgesamt und von N=22 (Saarland/2007) bis N=2826 (Nordrhein-Westfalen/1996) auf Bundeslandebene. Geringe Fallzahlen werden bei den folgenden Analysen unter entsprechendem Hinweis aus den Analysen ausgeschlossen, falls nötig. Eine Übersicht über die Fallzahlen pro Jahr für jeden Studiengang befindet sich im Anhang (Tab.A1).

[53] Für die Jahre 1996 und 1997 gibt es zwei Datenquellen, da die FDZ-Daten bereits ab 1996 vorliegen. In diesen Überschneidungsjahren beider Datenquellen ergeben sich zum Teil erhebliche Differenzen in den Fallzahlen. Dabei lagen die Angaben der amtlichen Daten in der Regel unter den Angaben aus den Archiven. Diese Differenzen wurden durch Angleichungen korrigiert, um konsistente Zeitreihen ohne künstliche Brüche zu erhalten.

Tabelle 7: Fallzahlen je Studiengang und Hochschule

		Göttingen	Braunschweig	Karlsruhe	Berlin	Tübingen	Heidelberg	Münster	Saarbrücken
Germanistik /Deutsch	Magister	1969-2010	XX	XX	1970-2010	1964-2010	1970-2010	1962-2010	1970-2010
	n=10 915	n=1 239			n=4 253	n=1 224	n=1 960	n=1 690	n=549
	Lehramt	1958-2010	1972-2010	1963-2010	1963-2010	1973-2010	XX	XX	XX
	n=17 950	n=4 504	n=1 065	n=4 502	n=4 547	n=3 332			
Mathematik	Diplom	1943-2010	1971-2010	1960-2010	1953-2010	1963-2010	1964-2010	1966-2010	XX
	n=9 471	n=1 473	n=790	n=1 421	n=1 248	n=863	n=1 283	n=2 393	
	Lehramt	1959-2010	1961-2010	1963-2010	1963-2010	1973-2010	XX	XX	XX
	n=6 744	n=1 172	n=847	n=2 245	n=1 254	n=1 226			
Chemie	Diplom	1943-2010	1970-2010	1960-2010	1950-2010	1970-2010	1959-2010	1950-2010	XX
	n=16 328	n=2 547	n=1 442	n=2 317	n=1 757	n=1 573	n=2 987	n=3 705	
Biologie	Diplom	1981-2010	1971-2010	1982-2010	1951-2010	1967-2010	1970-2010	1976-2010	XX
	n=19 190	n=3 878	n=1 751	n=906	n=3 455	n=4 138	n=2 977	n=2 085	
VWL	Diplom	1963-2010	XX	1960-2010	1953-2010	1950-2010	1955-2010	1950-2010	XX
	n=18 874	n=2 112		n=774	n=4 626	n=2 972	n=4 261	n=4 129	
BWL	Diplom	1963-2010	XX	1964-1981	1953-2010	1984-2010	XX	1957-2010	XX
	n=39 632	n=10 529		n=513	n=9 695	n=4 116		n=14 779	
Psychologie	Diplom	1967-2010	1972-2010	XX	1971-2010	1959-2010	1960-2010	1951-2010	XX
	n=18 860	n=1 745	n=1 567		n=5 606	n=3 171	n=3 104	n=3 667	
Soziologie	Magister/Diplom	1969-2010	XX	XX	1962-2010	1963-2010	1970-2010	1967-2010	XX
	n=6 721	n=552			n=4 177	n=523	n=619	n=850	
Maschinenbau	Diplom	XX	1972-2010	1960-2010	XX	XX	XX	XX	XX
	n=17 518		n=7 479	n=10 039					
Jura	Staatsexamen	Bundesweit 1959-2007 (1. SE: n=418 130 / 2. SE: n=347 031)							

Neben den nach Fach, Abschluss und Hochschule differenzierbaren Examensnoten umfassen die verfügbaren Daten Informationen zum Geschlecht der Prüflinge, zu den abgelegten Teilprüfungen und deren Ergebnissen, zu den jeweils geltenden Prüfungsordnungen und zu fachlichen Spezialisierungen. Die Vollständigkeit diese Informationen nimmt aufgrund starker Unterschiede in deren Verfügbarkeit in der Reihe ihrer Aufzählung ab. Für die Rechtswissenschaften sind neben den Prüfungsergebnissen Informationen zu (geschlechtsspezifischen) Durchfallquoten enthalten. Insgesamt sind die Prüfungen von 138 007 Prüflingen mit Abschluss Diplom, Magister oder Lehramt (1. Staatsexamen) bis 1997 auf dies Weise erfasst. Für 1998 bis 2010 liegen außerdem die Durchschnittsnotenwerte 140 473 weiterer bestandener Prüfungen vor. Zudem sind die Prüfungsergebnisse von 418 130 ersten und 347 031 zweiten Staatsexamina der Rechtswissenschaften (inklusive nicht bestandener Prüfungen) enthalten.

Die Erhebung der Daten erfolgte in den jeweiligen Universitätsarchiven sowie in einzelnen Fakultäts- und Institutsarchiven und, für die Fälle, in denen die relevanten Prüfungsakten noch nicht an das Archiv abgegeben wurden, in Prüfungsämtern. In der Regel wurden die Informationen zum Prüfungsergebnis entweder aus Zeugniskopien oder aus Prüfungsprotokollen aus den einzelnen Prüfungsakten der Prüflinge entnommen. In Einzelfällen, in denen keine Prüfungsakten überliefert wurden, konnten die Noten des entsprechenden Fachs außerdem aus Notenlisten erhoben werden. Die Abschlussnote wurde vor allem aus Gründen der mangelnden Verfügbarkeit differenzierterer Angaben stets in Prädikatsform erhoben, also als ganze Note von 1 (‚sehr gut') bis 4 (‚ausreichend'). Das in einigen Studiengängen an einigen Hochschulen vergebene Prädikat „ausgezeichnet" wurde der Vergleichbarkeit wegen als 1 (‚sehr gut') erfasst (vgl. Wissenschaftsrat 2012). In Jura wird bundesweit das Prädikat ‚voll befriedigend' vergeben, welches mit einer Gewichtung von 2.5 in die Berechnung der Durchschnittsnoten eingeht. An der FU Berlin wurde in den beiden wirtschaftswissenschaftlichen Studiengängen über einen kurzen Zeitraum ebenfalls das Prädikat ‚voll befriedigend' vergeben (in 87 Fällen), dies wurde in der Durchschnittsnotenberechnung ebenfalls mit dem Wert 2.5 gewichtet.

Die Teilprüfungsnoten wurden je nach Verfügbarkeit als ganze Noten oder als Dezimalnoten aufgenommen, für vergleichende Analysen wurden jedoch aus allen in differenzierter Form vorliegenden Teilprüfungsnoten zusätzlich standardisierte Noten in Form ganzer Noten gebildet, da von der Berichtsweise der Noten nicht immer darauf geschlossen werden kann, ob die Gesamtnote aus Teilprüfungsnoten in ganzer Form oder in Dezimalform berechnet wurde. Hier wurden Werte bis 1.50 einem ‚sehr gut', von 1.51 bis 2.50 einem ‚gut', von 2.51 bis 3.50 einem ‚befriedigend', von 3.51 bis 4.30 einem ‚ausreichend' und alle höheren Werte einem ‚nicht ausreichend' zugeordnet. Von der ursprünglich geplanten Berechnung von Durchfallquoten musste aufgrund diverser Probleme der Nachvollziehbarkeit nicht bestandener Prüfungen abgesehen werden (Gaens 2013).

Für einige Jahre konnten keine Prüfungsinformationen erhoben werden. Dies ist mehreren Gründen geschuldet: Für den Zeitraum vor 1997 sind vereinzelt keine vollständigen Dokumente im Archiv überliefert bzw. es wurden keine Prüfungen abgelegt. Im Zeitraum nach 1997 fehlen die Informationen für einzelne Jahre, da der Datensatz der amtlichen Prüfungsstatistik für diese Jahre Lücken aufweist oder die Fallzahlen zu gering sind, um die Freigabe der Daten zu erhalten. Wenn möglich und sinnvoll, wurden diese fehlenden Werte durch linear interpolierte Werte ersetzt[54].

Die Abschlussnoten für das Lehramt an den Baden-Württembergischen Hochschulen sowie für Diplom Chemie an der Universität Münster mussten über einige Jahre hinweg nachträglich berechnet werden, da im ersten Falle keine Gesamtabschlussnoten erteilt wurden, im zweiten Falle nur vereinzelt Prädikate überliefert wurden. Für das Lehramt wurde die Gesamtnote als arithmetisches Mittel der einfach gewichteten Teilprüfungsnoten (die Noten der jeweils belegten Fächer für das Lehramt) berechnet, wie es laut Auskunft im Landeslehrerprüfungsamt im entsprechenden Zeitraum auch für die Anfertigung interner Ranglisten bei Bewerbungen üblich war. Die nicht vermerkten Abschlussnoten der Chemieabsolvent*innen wurden nach den für den jeweiligen Zeitraum gültigen Berechnungsverfahren für die Gesamtnote berechnet, welche aus den Prüfungsordnungen und den Notenprotokollen entnommen werden konnten.
Die Prüfungsakten der Absolvent*innen im Lehramt werden in der Regel von den einzelnen Bundesländern oder von den untergeordneten Regierungsbezirken zentral für alle Landesprüfungsämter aufbewahrt. Durch fehlende Kennzeichnungen, an welcher Hochschule das Lehramtsstudium erfolgt ist, sind Differenzierungen nach Hochschulen in Berlin und für die Baden-Württembergischen Hochschulen nur begrenzt möglich. So enthalten die Daten für Berlin sowohl Absolvent*innen der Freien Universität Berlin als auch der Technischen Universität und seit 1990 auch der Humboldt-Universität. In den Daten für den Regierungsbezirk Karlsruhe sind neben den Karlsruher Absolvent*innen auch die aus Heidelberg enthalten, in den Daten für Tübingen auch die Absolvent*innen aus Ulm.

[54] Eine genaue Übersicht über die Jahre, für die keine Prüfungsinformationen vorliegen und über die ersetzten fehlenden Werte findet sich im Anhang (Tab.A2).

8. Ergebnisse

Der Ergebnisteil ist in einen Abschnitt zur deskriptiven Analyse und einen Abschnitt zur erklärenden Analyse unterteilt. Der erste Abschnitt zeichnet die Notengebung der berücksichtigten Studiengänge, zunächst für alle in den Daten enthaltenen Prüflinge auf Studiengangebene, dann an den in der Stichprobe enthaltenen Hochschulen nach[55]. Die deskriptive Analyse der Notenentwicklung erfolgt gemäß der perspektivischen Dualität von Quer- und Längsschnittbetrachtung in zwei Schritten. Die Betrachtung der Noten im (aggregierten) Querschnitt zielt vornehmlich auf die Beschreibung von Verhältnissen zwischen den Untersuchungseinheiten ab. Im Gegensatz zur eigentlichen Begriffsbedeutung wird im Folgenden häufig eine Betrachtung mehrerer Querschnitte gleichzeitig durchgeführt, da nicht einzelne Zeitpunkte im Fokus stehen, sondern mittel- und langfristig stabile Unterschiede und Gemeinsamkeiten. Einer Zusammenfassung der erklärungsbedürftigen Charakteristika folgt ein Abgleich der Daten mit bereits zuvor verfügbarem Datenmaterial, mit dem Ziel, die Generalisierbarkeit der Informationen auf Studiengangebene einzustufen.

Die Erklärung der im deskriptiven Abschnitt herausgearbeiteten Charakteristika der Notengebung im Zeitverlauf orientiert sich dann an den zuvor erarbeiteten Kategorien potentieller Ursachen für Unterschiede im Notenniveau und für die Entwicklung der Noten. Jeder potentielle Einflussfaktor wird dabei zunächst auf Studiengang-, dann auf Hochschullevel, jeweils erst im (aggregierten) Querschnitt, dann im Längsschnitt, betrachtet. Wann immer möglich, werden zur Analyse Daten herangezogen, die dem in Kapitel 7 beschriebenen Datensatz entstammen. Nicht im Datensatz enthaltene Informationen werden als Sekundärdaten in die Analysen eingebracht, sofern vorhanden und zugänglich.

8.1 Das Notenniveau an Hochschulen in Deutschland

8.1.1 Das Notenniveau in den untersuchten Studiengängen im Vergleich

Verteilungen der Abschlussnoten

Die Noten liegen bis 1997 als Individualdaten vor und lassen sich bis zu diesem Zeitpunkt vollständig darstellen. Über alle Jahre seit 1967, dem Jahr, ab dem für 11 der 12 Studiengänge Werte für mindestens zwei Hochschulen vorliegen, betrachtet, lassen sich aus den folgenden Histogrammen bereits einige Unterschiede zwischen den Studiengängen ablesen: Die zwischen 1967 und 1997 am häufigsten vergebene Examensnote ist in Mathematik Diplom, Biologie und Psychologie das Prädikat ‚sehr gut', die Verteilungen sind deutlich rechtsschief (Die Schiefe beträgt in Mathematik =0.886, in Biologie =1.033 und in Psychologie =0.748). In den beiden Lehramts- und den beiden Magisterstudiengängen, in Chemie und in Maschinenbau wurde in dieser Zeit das Prädikat ‚gut' am häufigsten

[55] Teile der folgenden Abschnitte zur deskriptiven Analyse wurden in veränderter Form bereits in Gaens (2015) veröffentlicht.

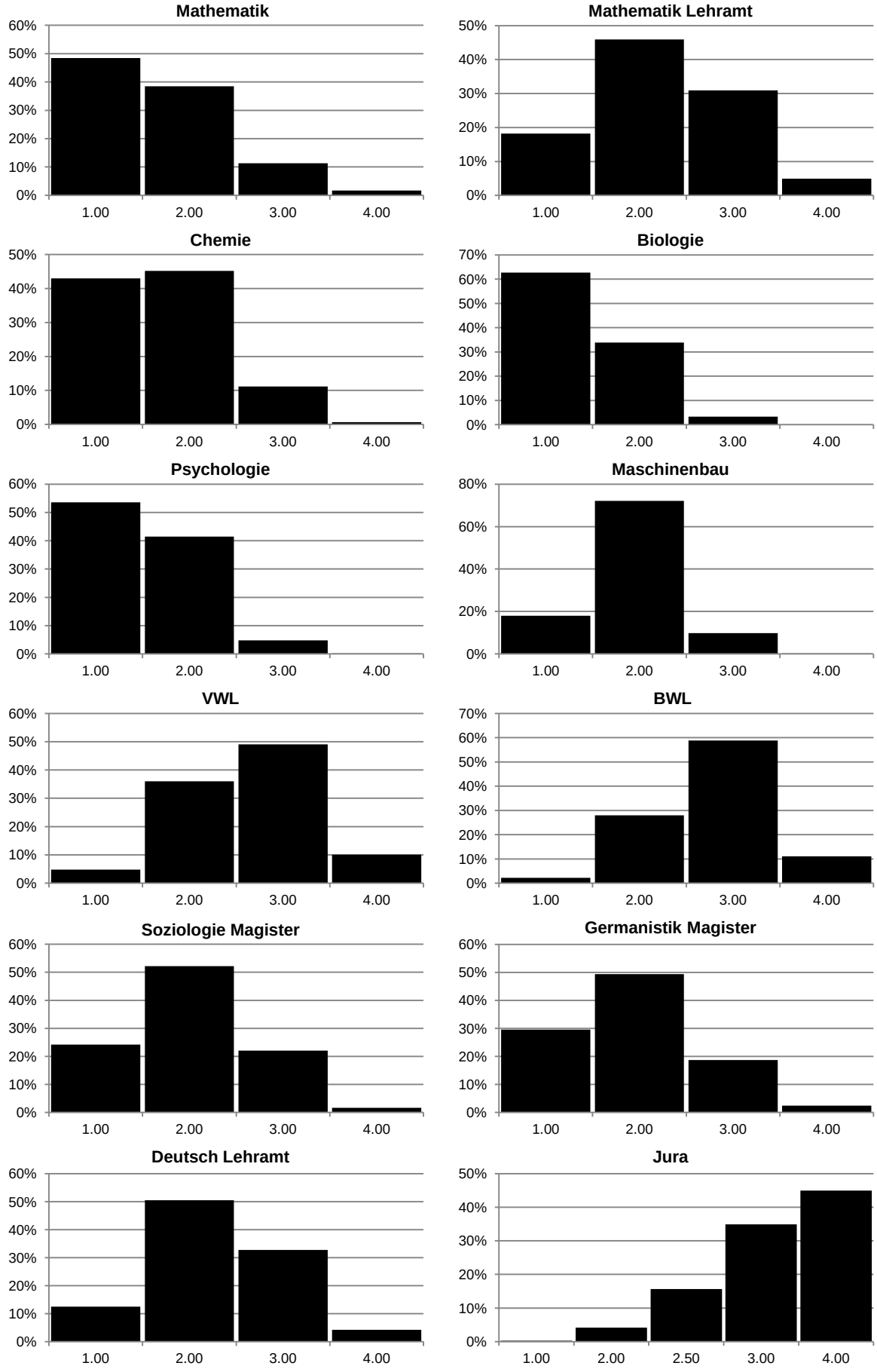
vergeben, in VWL und BWL das Prädikat ‚befriedigend'. Nur in Jura (das juristische ‚vollbefriedigend' ist mit dem Wert 2.50 übersetzt worden) ist ‚ausreichend' der Modus.

In VWL und den beiden Lehramtsstudiengängen kommt die Verteilung der Noten über alle Jahre hinweg am ehesten einer Normalverteilung nahe, wobei die deutliche Unimodalität nicht dem symmetrischen Charakter von Normalverteilungen entspricht. In den beiden Magisterstudiengängen und in Maschinenbau ist diese Symmetrie gegeben, allerdings sind die Noten hier (fast) nur über drei Ausprägungen verteilt: Von 6544 erfassten Abschlussnoten sind in Germanistik gerade einmal 156 ‚ausreichend' vergeben worden, was einen Anteil von 2.4% ausmacht. In Soziologie sind es 23 von 1411 Examensnoten, was 1.6% entspricht. In Maschinenbau ist der Anteil mit 0.2% noch einmal deutlich geringer (26 von 11752 Fällen), den gleichen Anteil nimmt die schlechteste Abschlussnote eines bestandenen Examens in Psychologie ein (23 von 11927 Fällen) - in Biologie sind es sogar nur 0.1% (15 von 11617 Fällen). Umgekehrt wurde in Biologie das Prädikat ‚sehr gut' in 62.7% der bestandenen Abschlussprüfungen vergeben, was den Spitzenwert unter den hier betrachteten Studiengängen darstellt. In BWL wurde diese Gesamtabschlussnote in 2.1%, in Jura sogar nur in 0.3% der Fälle vergeben - ein ‚ausreichend' gab es dort dagegen in 11.0% (BWL) bzw. 45.0% der Fälle.

Das vorhandene Beurteilungsspektrum von 1 (‚sehr gut') bis 4 (‚ausreichend') wird unterschiedlich stark genutzt - am breitesten in VWL, wo die Anteile der Extrempole 1 (4.8%) und 4 (10.0%) am ehesten ausgeglichen sind. Es folgen Deutsch Lehramt (‚sehr gut': 12.5%, ‚ausreichend': 4.2%) und Mathematik Lehramt (‚sehr gut': 18.2%, ‚ausreichend': 4.9%). In allen anderen Studiengängen liegt der Anteil an ‚ausreichend' bei maximal 2.4%, mit Ausnahme von BWL und Jura, wo der Anteil an ‚sehr gut' allerdings nur bei 2.1% bzw. 0.3% liegt. Am konzentriertesten ist die Nutzung der Notenskala in Biologie mit 96.5% ‚sehr guten' oder ‚guten' Abschlüssen bei 0.1% ‚ausreichend' und Psychologie mit 95.0% ‚sehr guten' oder ‚guten' Abschlüssen bei 0.2% ‚ausreichend'.

Über den Zeitraum 1967-1997 betrachtet, lassen sich anhand der Verteilungen der vergebenen Noten erste Unterschiede in der Notengebung zwischen den Studiengängen erkennen. Die Verteilungen zeigen, dass neben fachlichen Unterschieden auch Abschlussunterschiede bestehen, die es rechtfertigen, die Lehramtsnoten getrennt von den Noten in Diplom bzw. Magister zu analysieren. In Mathematik ist dieser Unterschied hinsichtlich der Verteilung größer als in Germanistik/Deutsch, aber auch dort ist er sichtbar. Da die Individualnoten nur bis 1997 vorliegen und die Verteilungen der Noten zudem über 31 Jahre betrachtet eventuelle zeitliche Veränderungen nicht abbilden, werden im Folgenden nun vor allem die Verteilungsparameter arithmetisches Mittel und Standardabweichung, die für alle Jahre bis 2010 vorliegen - wenn erforderlich, dann auch zeitabhängig - betrachtet.

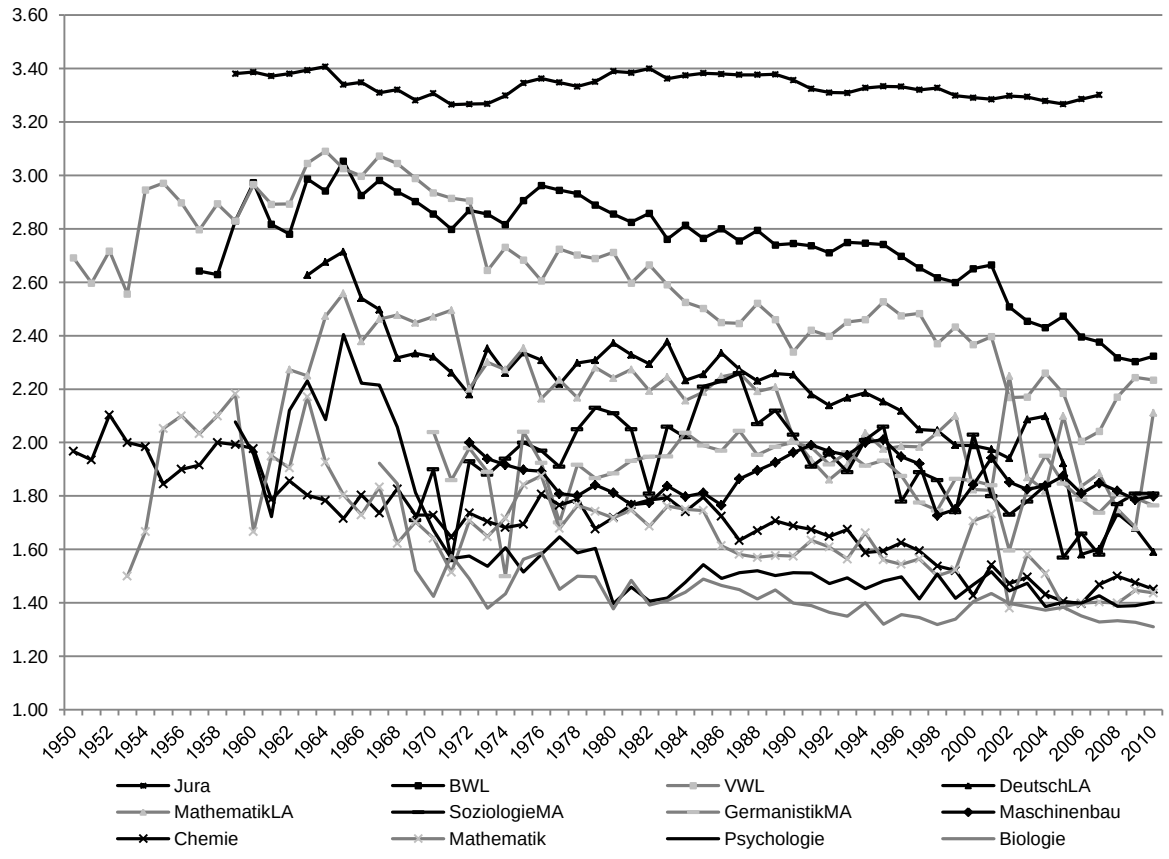
Abbildung 39: Verteilungen der Abschlussnoten in den Studiengängen von 1967-1997



Durchschnittliche Abschlussnoten

Abbildung 40 gibt einen ersten Überblick über die Entwicklung der durchschnittlichen Abschlussnoten auf Studiengangebene. Die im Folgenden dargestellten Studiengangdurchschnitte der Examensnoten stellen das arithmetische Mittel der Abschlussnoten aller Prüflinge eines Fachs mit gleichem Abschluss dar[56]. Es ist nur der Zeitraum abgebildet, für den für mindestens zwei Hochschulen Werte vorliegen, die Durchschnittsnoten in Jura sind über alle Hochschulen des gesamten Bundesgebiets gewichtet gemittelt. In Soziologie wurden nur an einer Hochschule (FU Berlin) über einen längeren Zeitraum Diplomprüfungen abgelegt (hier deshalb nicht abgebildet). Da in den Archiven keine verlässlichen Informationen über nicht bestandenen Prüfungen erhoben werden konnten und die amtliche Statistik nur endgültig nicht bestandene Prüfungen erfasst und damit ebenfalls keine aussagekräftigen Informationen enthält (Gaens 2013) geben die Durchschnittsnoten das Mittel der bestandenen Prüfungen wieder. Auf Studiengangebene gehen die Zeitreihen bis maximal 1950 zurück. In Soziologie (Magister) und Germanistik sind zu Beginn der Zeitreihen zwei bzw. sechs Datenpunkte mit geringen Fallzahlen (n≤13 bzw. n≤10) entfernt worden.

Abbildung 40: Die Entwicklung der Abschlussnoten in 12 Studiengängen im Zeitverlauf



[56] Die Gewichtung der Noten nach Studierendenzahl pro Hochschule erfolgt unter der Annahme, dass die Prüfungsergebnisse vorwiegend die tatsächlichen (über die Hochschulen zufällig verteilten) Leistungen der Studierenden abbilden und damit kein erhöhter Stichprobenfehler („Klumpeneffekt“) auftritt. Inwiefern diese Annahme empirisch haltbar ist, werden Analysen der Unterschiede im Notenniveau zwischen Hochschulen zeigen. Das juristische ‚vollbefriedigend‘ ist bei der Mittelwertbildung mit dem Wert 2.50 übersetzt worden, das an einzelnen Hochschulen in wenigen Jahren vergebene „Mit Auszeichnung“ wurde als ‚sehr gut‘ (1.0) kodiert.

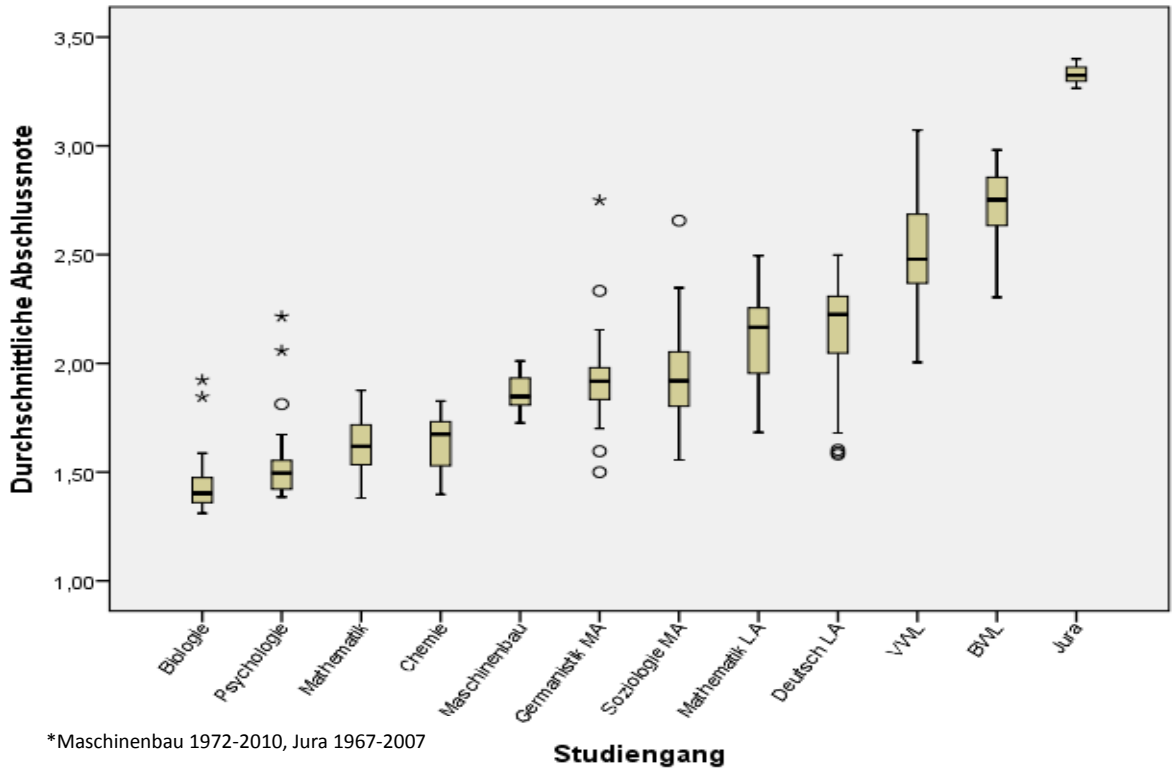
Die Grafik zeigt in Bezug auf die Notenhöhe, dass einerseits zu jedem erhobenen Zeitpunkt deutliche Unterschiede im gemittelten Notenniveau zwischen den meisten Studiengängen bestehen, es andererseits aber auch mehr oder weniger lange Perioden gibt, in denen sich einzelne Studiengänge in ihrem Notenniveau ähneln. Zudem zeichnet sich eine Notenhierarchie ab: In den juristischen Staatsprüfungen werden im Durchschnitt über den gesamten Zeitraum die schlechtesten Noten vergeben, es folgen BWL und VWL, wobei in VWL bis zum Beginn der 1970er Jahre schlechtere Noten vergeben werden als in BWL, bevor sich das Verhältnis umkehrt. Zwischen mehreren der übrigen acht Studiengänge finden immer wieder Überschneidungen im Notenniveau statt. Allerdings ist zu erkennen, dass in Psychologie und Biologie seit Beginn der 1970er Jahre stets die besten Noten vergeben werden, das Niveau in Chemie und Mathematik (Diplom) etwas höher ist und in Germanistik, Soziologie, Maschinenbau und den beiden Lehramtsstudiengängen die Noten spätestens ab 1990 nahe beieinander liegen, Letztere aber etwas über Ersteren liegen. Regressionen der Abschlussnoten auf Studiengangdummys für den Zeitraum von 1972 bis 2007, in dem für alle Studiengänge Werte vorliegen, zeigen, dass sich lediglich die Noten in Soziologie und Germanistik, Germanistik und Maschinenbau sowie Chemie und Mathematik nicht signifikant voneinander unterscheiden (siehe Anhang: Tab.A3-A14). Abbildung 41 und Tabelle 8 stellen die Unterschiede im Notenniveau noch deutlicher dar.

Abbildung 41 bildet die studiengangspezifische Streuung der Durchschnittsnoten über die Jahre ab: Die Boxplots zeigen jeweils die Verteilung der Durchschnittsnoten im Studiengang zwischen 1967, dem Jahr, ab dem für 11 der 12 Studiengänge Werte für mindestens zwei Hochschulen vorliegen und 2010. Lediglich in Maschinenbau wird die Verteilung der Noten über die Jahre erst ab 1972 illustriert, in Jura nur bis 2007. Die Abstufung der Studiengänge nach durchschnittlichem Notenniveau ist hier noch besser erkennbar. Eine Orientierung am Median (schwarzer Balken) zeigt, dass die Noten über die Zeit verteilt in Mathematik etwas besser als in Chemie und in Soziologie marginal besser als in Germanistik sind. Aufschlussreicher in Hinblick auf die langfristig stabilen Differenzen zwischen den Studiengängen sind jedoch die Spannweiten der jährlichen Durchschnittswerte: Die Box entspricht der Spannweite der 50% um den Median befindlichen Werte, die „Antennen“ enthalten die Jahre, die innerhalb des 1.5-fachen dieser Spannweite der 50% der mittleren Werte liegen, Kreise bzw. Sterne stellen die Jahre dar, die diese Spannweite mehr als 1.5 bzw. drei Mal überschreiten.

Während diese Spannweiten in Jura, Biologie, Psychologie und Maschinenbau am geringsten ausfallen, streuen die durchschnittlichen Examensnoten in Soziologie, VWL und den beiden Lehramtsstudiengängen (unter Nichtberücksichtigung der Ausreißerwerte) im Zeitverlauf am stärksten, das heißt in Ersteren ist das Notenniveau im Zeitverlauf am stabilsten, in Letzteren verändert es sich am stärksten. Die Überschneidungen der Boxen und Antennen zeigen an, dass im gesamten Zeitraum seit 1967 gleiche Notenniveaus vorkamen, was für alle Studiengänge außer für die Rechtswissenschaften zutrifft.

Dies ist nicht weiter verwunderlich, da, wie in Abbildung 40 ersichtlich, alle Studiengänge außer Jura im Zeitverlauf ihr Notenniveau nennenswert verändern und die gemeinsame Bandbreite an Werten damit von Jahr zu Jahr steigt. Überschneiden sich die Boxen und Antennen zweier Studiengänge nicht, bedeutet dies (wiederum unter Nichtberücksichtigung der Ausreißer), dass das durchschnittliche Notenniveau in den beiden Studiengängen im gesamten Zeitraum unterschiedlich hoch ist. So lässt sich aus der Grafik ablesen, dass das Notenniveau in Biologie seit 1967 wenn überhaupt, dann nur mit Ausreißerwerten an das Notenniveau heranreicht, das bestenfalls in Maschinenbau, Germanistik, in den beiden Lehramtsstudiengängen, in VWL, BWL und in Jura erreicht wird. Einen ‚eigenen' Bereich in der Notenskala hat jedoch nur Jura - in allen anderen Studiengängen existieren mehr oder weniger starke Überschneidungen des umspannten Wertebereichs mit dem anderer Studiengänge.

Abbildung 41: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1967-2010*



*Maschinenbau 1972-2010, Jura 1967-2007

Dennoch lässt sich anhand der Notenniveaus eine Rangfolge der Studiengänge erstellen, welche in Tabelle 8 ersichtlich ist. Dort sind die Studiengänge gemäß der über jeweils fünf Jahre gemittelten durchschnittlichen Notenhöhe für jeden dieser Fünfjahresabschnitte in steigender (also schlechter werdender) Reihenfolge aufgeführt. Aus den 12 Rängen lassen sich acht langfristig abgrenzbare Positionen in der Notenhierarchie bilden (Tab.8): Es bestätigt sich, dass in Biologie (1) die besten Noten vergeben werden, dicht gefolgt von Psychologie (2), in VWL (6), BWL (7) und mit Abstand in Jura (8), die schlechtesten. Mathematik und Chemie (3) lassen sich auch auf diese Weise von den Studiengängen mit den besten und schlechteren Noten abgrenzen - noch besser als in Abbildung 41 zeigt sich jedoch hier, dass Mathematik und Deutsch als Lehramtsstudiengänge ebenfalls eine eigene Position (5), zwischen BWL/ VWL auf der einen und Soziologie (Magister), Maschinenbau und Germanistik (4)

auf der anderen Seite, einzunehmen scheinen. Die Notenniveaus der Letzteren lassen sich untereinander nicht deutlich voneinander abgrenzen. Auffällig ist, dass die beiden Magister- ebenso wie beiden Lehramtsstudiengänge in ihrem Notenniveau nah beieinander liegen.

Tabelle 8: Rangfolge von Fünfjahresdurchschnitten der auf Studiengangebene gemittelten Durchschnittsnoten

	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	BIO	--	BIO	1.68	BIO	1.49	BIO	1.48	BIO	1.44	BIO	1.44	BIO	1.36	BIO	1.35	BIO	1.40	BIO	1.33
2	CHE	1.79	MAT	1.71	PSY	1.56	PSY	1.56	PSY	1.46	PSY	1.51	PSY	1.48	PSY	1.46	PSY	1.44	PSY	1.40
3	GER	1.94	CHE	1.76	MAT	1.69	CHE	1.75	MAT	1.74	MAT	1.58	MAT	1.61	CHE	1.54	CHE	1.47	MAT	1.42
4	MAT	1.95	PSY	2.00	CHE	1.69	MAT	1.76	CHE	1.78	CHE	1.68	CHE	1.64	MAT	1.57	MAT	1.52	CHE	1.46
5	PSY	2.11	SOZ	2.16	GER	1.85	MAB	1.83	MAB	1.80	MAB	1.88	GER	1.94	GER	1.82	SOZ	1.74	DLA	1.64
6	SOZ	--	GER	2.36	SOZ	1.86	GER	1.86	GER	1.97	GER	1.99	MLA	1.94	MAB	1.84	GER	1.81	SOZ	1.73
7	MAB	--	MAB	--	MAB	1.94	SOZ	2.03	SOZ	2.03	SOZ	2.14	SOZ	1.97	SOZ	1.86	MAB	1.87	GER	1.78
8	MLA	2.30	DLA	2.40	DLA	2.28	MLA	2.22	MLA	2.21	MLA	2.19	MAB	1.98	MLA	1.99	MLA	1.97	MAB	1.81
9	DLA	2.67	MLA	2.45	MLA	2.32	DLA	2.30	DLA	2.30	DLA	2.27	DLA	2.17	DLA	2.04	DLA	2.00	MLA	1.85
10	BWL	2.92	BWL	2.92	VWL	2.78	VWL	2.69	VWL	2.58	VWL	2.44	VWL	2.45	VWL	2.43	VWL	2.24	VWL	2.14
11	VWL	2.99	VWL	3.01	BWL	2.85	BWL	2.92	BWL	2.80	BWL	2.77	BWL	2.74	BWL	2.64	BWL	2.51	BWL	2.34
12	JUR	3.38	JUR	3.31	JUR	3.29	JUR	3.36	JUR	3.38	JUR	3.37	JUR	3.32	JUR	3.31	JUR	3.28	JUR	3.29

Es existieren also Unterschiede im Notenniveau, die es ermöglichen einzelne Studiengänge hinsichtlich der Bandbreite an Werten, innerhalb derer sich ihre durchschnittlichen Abschlussnoten im Zeitverlauf bewegen, voneinander abzugrenzen. Die Noten der 12 Studiengänge können zu jedem Zeitpunkt in eine Rangfolge gebracht werden, über Fünfjahresperioden gemittelt zeigen sich acht weitestgehend stabile Positionen, in die sie sich einteilen lassen. Wird das relativ stabile und stets schlechteste Notenniveau in Jura als Maßstab gesetzt und die Fünf-Jahres-Mittel der anderen Studiengänge als dessen prozentualer Anteil berechnet, ergibt sich folgendes Bild:

Tabelle 9: Rangfolge der Fünfjahresdurchschnitte in Prozent des Notenniveaus in Jura

	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	BIO	--	BIO	50.8	BIO	45.3	BIO	44.1	BIO	42.6	BIO	42.7	BIO	41.0	BIO	40.8	BIO	42.7	BIO	40.4
2	CHE	53.0	MAT	51.7	PSY	47.4	PSY	46.4	PSY	43.2	PSY	44.8	PSY	44.6	PSY	44.1	PSY	43.9	PSY	42.6
3	GER	57.4	CHE	53.2	MAT	51.4	CHE	52.1	MAT	51.5	MAT	46.9	MAT	48.5	CHE	46.5	CHE	44.8	MAT	43.2
4	MAT	57.7	PSY	60.4	CHE	51.4	MAT	52.4	CHE	52.7	CHE	49.9	CHE	49.4	MAT	47.4	MAT	46.3	CHE	44.4
5	PSY	62.4	SOZ	65.3	GER	56.2	MAB	54.5	MAB	53.3	MAB	55.8	GER	58.4	GER	55.0	SOZ	53.1	DLA	49.9
6	SOZ	--	GER	71.3	SOZ	56.5	GER	55.4	GER	58.3	GER	59.1	MLA	58.4	MAB	55.6	GER	55.2	SOZ	52.6
7	MAB	--	MAB	--	MAB	59.0	SOZ	60.4	SOZ	60.1	SOZ	63.5	SOZ	59.3	SOZ	56.2	MAB	57.0	GER	54.1
8	MLA	68.1	DLA	72.5	DLA	69.3	MLA	66.1	MLA	65.4	MLA	65.0	MAB	59.6	MLA	60.1	MLA	60.1	MAB	55.0
9	DLA	79.0	MLA	74.0	MLA	70.5	DLA	68.5	DLA	68.1	DLA	67.4	DLA	65.4	DLA	61.6	DLA	61.0	MLA	56.2
10	BWL	86.4	BWL	88.2	VWL	84.5	VWL	80.1	VWL	76.3	VWL	72.4	VWL	73.8	VWL	73.4	VWL	68.3	VWL	65.1
11	VWL	88.5	VWL	90.9	BWL	86.6	BWL	86.9	BWL	82.8	BWL	82.2	BWL	82.5	BWL	79.8	BWL	76.5	BWL	71.1
12	JUR	100.0	JUR	100.0	JUR	100.0	JUR	100.0	JUR	100.0	JUR	100.0	JUR	100.0	JUR	100.0	JUR	100.0	JUR	100.0

In Biologie beträgt das Notenniveau seit 1971 ca. 40% bis 45% des Notenniveaus in Jura, ist also mehr als doppelt so gut. Genauer gesagt, beträgt die Differenz etwa 55 bis 60 Prozentpunkte. Psychologie liegt mit 43% bis 47% des Juraniveaus im selben Zeitraum im Überschneidungsbereich dieser Werte. Mathematik (43% bis 52% des Juraniveaus) und Chemie (44% bis 53%) verringern die prozentuale

Differenz zu den Juranoten weiter und befinden sich dabei beide im gleichen Bereich, immer noch mit leichter Überschneidung zur Spannweite in Biologie und Psychologie. Germanistik (54% bis 59%) und Maschinenbau (53% bis 60%) grenzen sich schon deutlicher in ihrer relativen Stellung ab. Das relative Notenniveau in Soziologie (53% bis 64%) bewegt sich in dieser Betrachtung als Grenze zwischen Germanistik und Maschinenbau auf der einen und Deutsch Lehramt (61% bis 69%, 50% im letzten Fünf-Jahres-Abschnitt) sowie Mathematik Lehramt (56% bis 71%) auf der anderen Seite. In VWL (65% bis 85%) sind auch noch Überschneidungen mit der Spannweite in den Lehramtsstudiengängen ersichtlich, in BWL (71% bis 87%) nur noch mit VWL. Der Vollständigkeit halber seien auch noch die über den gesamten Zeitraum seit 1967 gemittelten Noten inklusive der prozentualen Anteile am Juraniveau angegeben:

Tabelle 10: Von 1967-2010 ungewichtet gemittelte Durchschnittsnoten und relativer Anteil am Juraniveau

Studiengang	Mittelwert Noten seit 1967	Anteil am Notendurchschnitt Jura
Biologie	1.44	43.2 %
Psychologie	1.53	45.9 %
Mathematik	1.62	48.6 %
Chemie	1.64	49.2 %
Maschinenbau (ab 1972)	1.87	56.2 %
Germanistik Magister	1.92	57.7 %
Soziologie Magister	1.94	58.3 %
Mathematik Lehramt	2.12	63.7 %
Deutsch Lehramt	2.15	64.6 %
VWL	2.52	75.7 %
BWL	2.72	81.7 %
Jura	3.33	100.0 %

Auch hier zeigt sich, dass sich a) Biologie, b) Psychologie, c) Mathematik und Chemie, d) Maschinenbau, Germanistik und Soziologie, e) Mathematik Lehramt und Deutsch Lehramt, f) VWL, g) BWL und h) Jura anhand der prozentualen Anteile gruppieren lassen, wie es bereits zuvor geschehen ist.

Da die bis hier erfolgten Einordnungen alle auf der Basis des durchschnittlichen Notenniveaus vorgenommen wurden, sollte geklärt sein, wie aussagekräftig das genutzte Lagemaß des arithmetischen Mittels für die Notengebung in den einzelnen Studiengängen überhaupt ist. Auskunft darüber gibt die jeweilige Streuung der Noten um den Mittelwert. Da die Streuung für jeden einzelnen Jahreswert der Zeitreihe im Vergleich der 12 Studiengänge schwer zu vergleichen ist, wird jeweils das Mittel der Standardabweichungen über alle Jahre berechnet und dessen Standardabweichung und damit die durchschnittliche Abweichung von der Abweichung genutzt, um die Aussagekraft dieser Mittelung zu bewerten.

Tabelle 11: Streuung der Noten in den einzelnen Studiengängen seit 1967

Studiengang	Mittlere Standardabweichung	Standardabweichung der Standardabweichung
Maschinenbau	0.52	0.05
Biologie	0.57	0.07
Psychologie	0.59	0.06
BWL	0.64	0.06
Jura	0.66	0.02
Chemie	0.67	0.06
VWL	0.69	0.05
Deutsch Lehramt	0.71	0.05
Mathematik	0.71	0.08
Soziologie Magister	0.72	0.10
Mathematik Lehramt	0.75	0.09
Germanistik Magister	0.75	0.07

Die Standardabweichung der Standardabweichung zeigt, dass sich in den meisten Studiengängen eine ähnliche Streuung der Standardabweichungen über die Zeit ergibt. Die über die Jahre gemittelte Standardabweichung kann damit als aussagekräftiger Vergleichswert bezüglich der Notenstreuung in den einzelnen Studiengängen eingestuft werden.

Die Streuung der Noten selbst erstreckt sich von durchschnittlich s=0.52 Noten in Maschinenbau bis s=0.75 in Germanistik Magister und Mathematik Lehramt. Diese Werte zeigen, dass die Durchschnittsnote als aussagekräftig für die Verteilung der Noten in den einzelnen Studiengängen angesehen werden kann: Bei einer Notenskala von 1 bis 4, aus der nur ganze Zahlen vorliegen beträgt die maximal zu erreichende Streuung s=1.5, wenn genau die Hälfte aller Studierenden eine 4, die andere Hälfte eine 1 erhalten. Bei einer gleichmäßigen Verteilung nur auf die Noten 1 und 3 bzw. 2 und 4 wäre s=1, teilen sich die Noten nur auf zwei Noten auf, die nebeneinander liegen, ist s=0.5. Ein Wert von ca. s=0.7 kommt beispielsweise durch eine Verteilung zustande, in der jeweils ein Viertel der Werte eine Note über und ein Viertel eine Note unter der Note liegen, die mit 50% der Werte den Modus darstellt. Die empirischen Werte, die zwischen einer halben bis dreiviertel Note Abweichung vom Mittelwert liegen, zeugen damit von einer geringen Streuung.

Sie zeigen, dass die Breite der Notenskala in Maschinenbau am geringsten, in Mathematik Lehramt und Germanistik Magister am stärksten ausgenutzt wird. Bemerkenswert ist, dass es scheinbar keinen generellen Zusammenhang zwischen dem Notenniveau und dem Ausnutzen der Notenskala gibt: Maschinenbau und BWL weisen als Studiengänge mit vergleichsweise schlechtem Durchschnittsnoten eine geringere Streuung auf als etwa Mathematik, wo die Noten im Mittel wesentlich besser sind. Umgekehrt ist es nicht weiter verwunderlich, dass Jura auf der einen und Biologie und Psychologie auf der anderen Seite sehr homogene Notenverteilungen aufweisen. Hier fungiert aufgrund des extrem schlechten bzw. extrem guten Notenniveaus das obere bzw. untere Ende der Notenskala als Begrenzung für die Streuung. Ein sehr schlechtes oder gutes Notenniveaus geht demnach mit einer geringen Streuung einher, während in der Mitte der Notenskala keine Kopplung zwischen Lage- und Streuungsmaß ersichtlich ist.

Auch wenn anhand des Vergleichs der absoluten und relativen Notenniveaus der einzelnen Studiengänge bereits erklärungsbedürftige Differenzen aufgezeigt werden können, geht aus deren alleiniger Betrachtung noch nicht hervor, wie bedeutsam eigentlich eine bestimmte Differenz x1 zwischen zwei Studiengängen im Vergleich zu einer anderen Differenz x2 zwischen zwei anderen Studiengängen ist. Um dies zu klären, wird im Folgenden ein Versuch unternommen, die Unterschiede im Notenniveau zwischen den Studiengängen zu systematisieren.

In einem ersten Schritt werden dazu Varianzanalysen durchgeführt, für jedes einzelne Jahr seit 1967. Nicht weiter überraschend ergibt bereits die einfaktorielle ANOVA (**AN**alysis **O**f **VA**riance) für jedes Jahr bis 1997 signifikante Unterschiede zwischen den Abschlussnoten der einzelnen Studiengänge[57] (siehe Anhang: Tab.A15). Dies besagt jedoch nur, dass sich zu jedem Messzeitpunkt mindestens zwei Studiengänge signifikant im Notenniveau unterscheiden. Da in keinem Jahr Varianzhomogenität vorliegt, ist zudem eine der Anwendungsvoraussetzung für eine ANOVA nicht erfüllt. Detailliertere Erkenntnisse bietet ein Post-Hoc Test, der einzelne Paarvergleiche zwischen den Faktorgruppen durchführt. Aufgrund der Varianzungleichheit und seiner Robustheit gegenüber unterschiedlich großen Fallzahlen wird der Games-Howell Test gewählt (Field 2013:374f).

[57] Zur groben Einschätzung, bei welchem Ausmaß die Unterschiede im Notenniveau statistisch signifikant ($p \leq 0.05$) sind, kann dem Leser folgender Orientierungswert dienen: Bei einer Fallzahl von n=50 und einer Standardabweichung von $\sigma=0.5$ unterscheiden sich zwei Durchschnittsnoten $\bar{x}1$ und $\bar{x}2$ zwischen zwei Studiengängen ab einer Differenz $\bar{x}1 - \bar{x}2$ von ca. 0.14 ($= 1.96*0.5/\sqrt{50}$, $\alpha = 0.05$) signifikant voneinander.

Tabelle 12: Zeiträume/Zeitpunkte signifikant differenter Notenniveaus zwischen den Studiengängen (Paarvergleiche)

	Biologie	Psychologie	Chemie	Mathematik	Maschinenbau	GermanistikMA	SoziologieMA	MathematikLA	DeutschLA	VWL	BWL	Jura
Jura	1967-2007	1967-2007	1967-2007	1967-2007	1972-2007	1969-2007	1969-2007	1967-2007	1967-2007	1967-2007	1967-2007	--
BWL	1967-2010	1967-2010	1967-2010	1967-2010	1972-2010	1969-2010	1969-2010	1967-2001 2003-2004 2008-2009	1967-2010	1973 1975-1979 1981-2008	--	
VWL	1967-2010	1967-2010	1967-2010	1967-2010	1972-2010	1969-2010	1969-1984 1988-1989 1991-2010	1967-1985 1991-1997 2000-2001 2008-2009	1967-1985 1988-1989 1991-2001 2005-2010	--		
DeutschLA	1969-2010	1969-2009	1967-2005 2008-2009	1967-1999 2002-2005 2007-2009	1973-1991 1993-1999 2003 2006-2007	1971 1973-1974 1976-1990 1992-1998 2002- 2003 2006	1971/1973 1977/1996 2005	xx	--			
MathematikLA	1969-1999 2002/2005 2008-2009	1968-1999 2002/2005	1967-1989 1994-1999 2002/2005	1967-1989 1993-1999 2002/2005	1973-1987	1971 1973-1974 1977 1979-1983 1986/1997 2002	1971/1977	--				
SoziologieMA	1973-1975 1977-1981 1983-2000 2003-2004 2006-2010	1975 1977-1981 1983-1995 1997-2000 2003-2004 2006 2008-2010	1979 1985-1989 1994-1995 2000 2003-2004 2006 2008-2010	1985-1990 1994-1995 1997-1998 2006 2008-2010	1986-1987 2005/2007	2005	--					
GermanistikMA	1970 1972-1973 1975 1978-2010	1975-1976 1978-2001 2003-2010	1984 1986-2001 2003-2010	1982/1984 1986-1999 2002-2010	1974/1984 1986-1987 1997/2002	--						
Maschinenbau	1972-2010	1972-2010	1972-1975 1987-2010	1972-1973 1987-1999 2002-2010	--							
Mathematik	1973-1985 1990-2001	1968 1975-1976 1980-1984 1994/1997 2001	2000	--								
Chemie	1973-1974 1976-1978 1980-1999 2008-2009	1967 1975-1976 1978 1980-1986 1988-1994 1997	--									
Psychologie	1977 1990-1993 1995-1996 1998	--										
Biologie	--											

Die Ergebnisse sind in Tabelle 12 festgehalten. Sie gibt einen Überblick über alle Jahre, in denen sich das Notenniveau zwischen den einzelnen Studiengängen signifikant ($p \leq 0.05$) unterscheidet. Die Anzahl der signifikant differenten Jahre pro Paarvergleich gibt an, ob die zuvor in der Rangfolge verdeutlichten Unterschiede im Notenniveau auch als relevantes Unterscheidungskriterium in der Notengebungspraxis zweier Studiengänge herangezogen werden können, oder ob sie so gering sind, dass sie trotz einer kontinuierlichen Rangfolge nicht zur Differenzierung geeignet sind. Um die Verhältnisse zwischen Paarvergleichen, welche mit und ohne Beteiligung von Jura und Maschinenbau (drei bzw. fünf Jahreswerte weniger) durchgeführt wurden, einstufen zu können, wurde für jeden Paarvergleich ein relationaler Wert (Anzahl signifikant differente Jahre dividiert durch die Gesamtzahl an Jahren) berechnet. Anhand dieses relationalen Wertes wurden die Studiengänge in fünf Klassen gleicher Breite (je 20% der Werteskala) eingeteilt und für jede Klasse wieder die Werte der Jahre (zurück)berechnet, die sich über diese Klasse verteilen. Es ergeben sich durch die Klassierung folgende (wieder in Jahre zurückgerechnete) Wertebereiche in den Daten:

Klasse	Anzahl signifikant differenter Werte
1 keine/kaum signifikante Differenzen (n=8)	0 - 8 Jahreswerte (Maschinenbau (Mb):4-6)
2 gelegentlich signifikante Differenzen (n=5)	11-15 Jahreswerte (Mb: 15)
3 ausgeglichen signifikant/nicht signifikant (n=4)	19-25 Jahreswerte
4 häufig signifikante Differenzen (n=14)	27-35 Jahreswerte (Mb: 24-29)
5 durchgängig signifikante Differenzen (n=35)	37-44 Jahreswerte (Mb: 39; Jura 39-41; Jura/Mb:36)

Tabelle 13 gibt diese Beziehungsklassen für jeden Paarvergleich aus. Grundsätzlich zeigt sich ein zu erwartendes Muster: Je näher die Studiengänge in der zuvor anhand der Fünfjahresmittelwerte gebildeten Rangfolge aneinander liegen, umso niedriger wird die Beziehungsklasse, das heißt, umso seltener werden die Jahre, in denen sie sich signifikant hinsichtlich ihres Notenniveaus unterscheiden. Die Häufigkeit der signifikanten Jahreswerte sinkt dabei nicht proportional zur Rangfolge, sondern steht in Bezug zur Höhe der durchschnittlichen Differenz zwischen den Notenniveaus im Zeitverlauf.

Die Klassierung weist darauf hin, dass es irreführend wäre, Biologie und Psychologie als unterschiedliche Positionen aufzufassen, wie es die einfache Rangfolge nahelegt: Nur acht von 44 Jahreswerten weisen einen signifikanten Unterschied auf. Auch die Zuordnung von Soziologie in eine Gruppe mit Maschinenbau und Germanistik ist nicht so eindeutig, wie es die Rangfolge der absoluten Noten nahelegt. Vielmehr scheint Soziologie nicht eindeutig in eine Gruppe zu fallen, da das Notenniveau auch gegenüber den beiden Lehramtsstudiengängen eine ähnlich geringe Anzahl von Jahreswerten aufweist wie gegenüber Maschinenbau und Germanistik auf der anderen Seite. Die hohen Stufenwerte für BWL, VWL und Mathematik/Chemie bestätigen hingegen deren Abgrenzung gegenüber den anderen Studiengängen als eigenständigen Positionen.

Anzumerken bleibt, dass trotz gemeinsamer Positionierung mehrerer Studiengänge aufgrund niedriger Abgrenzungswerte untereinander, unterschiedliche Beziehungsstufen zu den Studiengängen auf den anderen Positionen bestehen. So weist Psychologie einem geringeren Abstand zu Chemie, Mathematik und Germanistik auf als Biologie, was aufgrund des durchgängig niedrigeren durchschnittlichen Notenniveaus (wenn auch nicht signifikant niedriger) nicht weiter überraschend ist. Hilfreicher ist die Abstufung der Studiengänge nach den Beziehungsklassen, wenn Beziehungen zwischen den Studiengängen betrachtet werden, welche nicht konstant auf dem gleichen Rang verbleiben. So zeigt sich, dass Mathematik und Chemie seltener signifikante Differenzen gegenüber Soziologie als gegenüber Germanistik und noch seltener gegenüber Maschinenbau aufweisen. Zu Letzterem besteht ein ähnliches Verhältnis wie zu Mathematik Lehramt, das in der Rangfolge hinter Maschinenbau steht, jedoch eine Position mit Soziologie bildet.

Eine Überprüfung der überarbeiteten Positionsbildung
1: Biologie/Psychologie
2: Mathematik/Chemie
3: Maschinenbau/Germanistik/Soziologie
4: Soziologie/Mathematik Lehramt/Deutsch Lehramt
5: VWL
6: BWL
7: Jura
über die Begrenzung auf den Zeitraum 1967 bis 2010 hinaus ist für die jeweils aneinander grenzenden Studiengänge nur für Vergleiche mit den Studiengängen Psychologie (ab 1959), Mathematik/Chemie (ab 1950), Germanistik (ab 1964), Mathematik Lehramt (ab 1961), Deutsch Lehramt (ab 1963), VWL (ab 1950), BWL (ab 1957) und Jura (ab 1959) möglich. Für Soziologie, Germanistik und Biologie (ab 1967) sowie Maschinenbau (ab 1972) liegen keine früheren Daten vor.
Die (im Vergleich zu den Post-Hoc Ergebnissen etwas liberaleren) Ergebnisse der möglichen T-Tests für die aneinandergrenzenden Paare vor 1967 zeigen jedoch nur durchgängig signifikante Unterschiede für die Grenzen Jura/BWL (acht von acht Jahren vor 1967), VWL/Mathematik Lehramt (6/6) und VWL/Deutsch Lehramt (4/4). Zwischen Chemie und Psychologie erweisen sich vier der acht zusätzlich verfügbaren Vergleichsjahre als signifikant different, zwischen Mathematik und Psychologie sind es nur zwei von acht, zwischen BWL und VWL zwei von 10 und zwischen Deutsch Lehramt und Germanistik, Germanistik/Mathematik sowie Germanistik/Chemie jeweils eines von drei. Die Grenzen Mathematik Lehramt/Germanistik weisen keinen zusätzlichen signifikanten Wert auf. Diese Ergebnisse passen mit zwei Ausnahmen in allen Vergleichsfällen zur sich ab 1967 anschließenden Struktur. Das heißt, ergeben die Post-Hoc Vergleiche (ab) 1967 eine signifikante Differenz, sind auch die

Jahre für die vor 1967 Werte vorliegen signifikant different und umgekehrt. Lediglich zwischen Mathematik/Psychologie und Chemie/Psychologie sind die Werte 1968 bzw. 1967 signifikant, vorher jedoch nur in zwei bzw. vier von acht Jahren. Konkret bedeutet dies, dass die Grenzen zwischen Jura/BWL, VWL/Deutsch Lehramt und VWL/Mathematik Lehramt, also zwischen den Positionen 7 und 6 sowie 5 und 4 auch schon vor 1967 Bestand hatten. Bedingt gilt dies auch für die Grenze zwischen Position 1 und 2 (Chemie und Mathematik/Psychologie und Biologie). Für die Abgrenzungen zwischen BWL und VWL (Position 6 und 5) sowie zwischen Deutsch Lehramt/Germanistik bzw. Mathematik Lehramt/Germanistik (Position 4 und 3) auf der einen und Germanistik/Mathematik bzw. Germanistik/Chemie (Position 3 und 2) auf der anderen Seite kann hingegen festgehalten werden, dass sich in den 1960er Jahren noch keine nennenswerten Unterschiede im Notenniveau finden lassen.

Tabelle 14 enthält pro Zelle drei Angaben: Im oberen Zellbereich den Zeitraum bzw. die Zeiträume mit *durchgängig* signifikanter Differenz zwischen den jährlichen Durchschnittswerten. Als durchgängig sind hier Zeiträume definiert, die aus mindestens drei aufeinanderfolgenden Zeitpunkten bestehen und innerhalb derer nur Lücken von maximal einem nicht signifikanten Wert zwischen zwei signifikanten Werten auftreten. Enthält der abgebildete Zeitraum solche einjährige Lücken ist dies, gemeinsam mit der Anzahl solcher Lücken, vermerkt. In den unteren Zellbereichen befinden sich auf der linken Seite des Trennstrichs Angaben zur Dauer der in den oberen Zellbereichen dargestellten Zeiträume durchgängig signifikanter Unterschiede. Rechts vom Trennstrich ist angegeben, wie viele Jahre seit 1967 insgesamt, unabhängig von ihrer zeitlichen Abfolge, eine signifikante Differenz im Notenniveau aufweisen.
Während die Häufigkeiten der Unterscheidungswerte grundsätzliche Aussagen über die Differenzierung der einzelnen Notenniveaus im Hinblick auf den gesamten betrachteten Zeitraum ermöglichen, erlauben sie noch keine Aussagen über die zeitliche Ausgestaltung dieser Differenzierungslinien. Von Bedeutung für die Erklärung von *stabilen* Differenzen sind Erkenntnisse über die zeitliche Dimension, einmal in Bezug auf die Dauer *durchgängiger* Unterschiede und einmal in Bezug auf den konkreten Zeitraum, in dem sie existieren. Auch die Dauer der durchgängigen Unterscheidungsperioden (insgesamt 81 Perioden bei 66 Paarvergleichen) ist in wiederum fünf gleich breiten Klassen angegeben. Um die Vergleichbarkeit zwischen den Paarvergleichen unter Beteiligung von Maschinenbau mit den übrigen zu gewährleisten wurde auch hier durch Division der ermittelten Periodendauern durch die maximal erreichbare Dauer im jeweiligen Vergleichsrahmen für alle Paarvergleiche ein relationaler Wert berechnet. Die Klassierung der Werte ergibt folgende zeitliche Abstufungen:

Klasse	Dauer durchgängig signifikanter Perioden
1 kein stabiler Unterschied (n=23)	0 - 8 Jahre (Maschinenbau (Mb): 0-3)
2 kurzfristig stabile Unterschiede (n=8)	11-17 Jahre (Mb: 8-14)
3 mittelfristig stabile Unterschiede (n=7)	18-25 Jahre
4 längerfristig stabile Unterschiede (n=10)	26-32 Jahre (Mb: 23-26)
5 langfristig stabiler Unterschied (n=33)	35-43 Jahre (Mb: 38; Jura 38-40; Jura/Mb: 35)

Tabelle 13: Beziehungsklassen der Notenniveaus zwischen den einzelnen Studiengängen

	Biologie	Psychologie	Chemie	Mathematik	Maschinenbau	GermanistikMA	SoziologieMA	MathematikLA	DeutschLA	VWL	BWL	Jura
Jura	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	--
BWL	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	4 | 5	--	5 | 5
VWL	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 5	5 | 2;3	4 | 1;3	5 | 1;2;3	--	4 | 5	5 | 5
DeutschLA	5 | 5	5 | 5	5 | 5	5 | 1;4	4 | 4	4 | 4	1 | 1	1 | 1	--	5 | 1;2;3	5 | 5	5 | 5
MathematikLA	4 | 4	4 | 4	4 | 1;3	4 | 1;3	2 | 2	2 | 1	1 | 1	--	1 | 1	4 | 1;3	5 | 5	5 | 5
SoziologieMA	4 | 1;4	4 | 1;3	2 | 1;1	2 | 1;1	1 | 1	1 | 1	--	1 | 1	1 | 1	5 | 2;3	5 | 5	5 | 5
GermanistikMA	5 | 4	4 | 5	3 | 4	3 | 1;2	1 | 1	--	1 | 1	2 | 1	4 | 4	5 | 5	5 | 5	5 | 5
Maschinenbau	5 | 5	5 | 5	4 | 1;4	4 | 2;2	--	1 | 1	1 | 1	2 | 2	4 | 4	5 | 5	5 | 5	5 | 5
Mathematik	3 | 2;2	2 | 1	1 | 1	--	4 | 2;2	3 | 1;2	2 | 1;1	4 | 1;3	5 | 1;4	5 | 5	5 | 5	5 | 5
Chemie	4 | 4	3 | 3	--	1 | 1	4 | 1;4	3 | 4	2 | 1;1	4 | 1;3	5 | 5	5 | 5	5 | 5	5 | 5
Psychologie	1 | 1	--	3 | 3	2 | 1	5 | 5	4 | 5	4 | 1;3	4 | 4	5 | 5	5 | 5	5 | 5	5 | 5
Biologie	--	1 | 1	4 | 4	3 | 2;2	5 | 5	5 | 4	4 | 1;4	4 | 4	5 | 5	5 | 5	5 | 5	5 | 5

a | b bzw. a | bi ist zu lesen als: Beziehungsklasse nach Gesamtzahl signifikant differenter Jahreswerte | Beziehungsklassen des Zeitraums (b); der Zeiträume (b; bi) mit durchgängig signifikantem Unterschied im Notenniveau

Der Vergleich der Beziehungsklassen der Gesamtzahl signifikant differenter Werte mit den Dauern der Perioden durchgängig signifikanter Differenzen zeigt, dass die Positionsbildung nach ersteren nicht ausreicht, um Aussagen über die zeitliche Stabilität dieser Unterschiede zu treffen. So zeigt sich in den Abgrenzungen Psychologie vs. Mathematik, Mathematik/Chemie vs. Soziologie und Germanistik vs. Mathematik Lehramt, dass die Abgrenzung über die Positionsrangfolge vor allem durch mehrere kurze Perioden oder Einzeljahre, in denen sich die Noten signifikant unterscheiden, zustande kommt. Die zeitliche Stabilität der Unterschiede ist in diesen Fällen aber deutlich geringer einzustufen.

Einen Überblick über die zeitliche Stabilität der vorgenommenen Positionsabgrenzungen bietet Abbildung 42. Es wird sichtbar, dass die Positionsabgrenzung mit Ausnahme des Extremfalls Jura immer unter der Berücksichtigung der zeitlichen Dimension verstanden werden sollte. So beruht die Abgrenzung von Biologie und Psychologie gegenüber Mathematik und Chemie ebenso wie die von Germanistik und Maschinenbau gegenüber Deutsch Lehramt vor allem auf den Notendifferenzen, die zwischen dem Beginn der 1970er und Ende der 1990er/ Beginn der 2000er Jahre auftreten. Gegenüber Mathematik Lehramt bestehen die Differenzen in Germanistik und Maschinenbau sogar nur bis

in die 1980er Jahre. Chemie und Mathematik lassen sich gegenüber Germanistik, Maschinenbau und Soziologie hingegen zwar erst seit den 1980ern klar abgrenzen, dafür hält diese Abgrenzung aber bis zum letzten Messzeitpunkt 2010 an. VWL setzt sich zwar fast über den gesamten Zeitraum in Richtung besserer Noten von den Lehramtsstudiengängen ab, aber auch hier ist zu beachten, dass sich diese Differenzen gegenüber Deutsch Lehramt Mitte der 1980er und Anfang der 2000er Jahre, gegenüber Mathematik Lehramt auch Mitte der 1980er und danach immer wieder kurzzeitig, aufheben. Ebenso beachtenswert ist, dass Differenzen gegenüber BWL, von dem sich VWL in Richtung schlechterer Noten abhebt, erst zu Beginn der 1970er entstehen, dann lange anhalten, zum Ende der 2000er Jahre dann aber wieder verschwinden.

Entsprechend der höheren Konzentration der Studiengänge in der besseren Hälfte der Notenskala steigen die durchschnittlichen Distanzen im Notenniveau zwischen den Positionen zunehmend an. Zwischen den Positionen 1 und 2 liegt die gemittelte Betragsfunktion[58] der jährlichen Differenzen (im Folgenden zur Unterscheidung von der tatsächlichen mathematischen Differenz als Abstand oder Distanz bezeichnet) im Bereich von 0.23 bis 0.30 Noten innerhalb der Perioden durchgängig signifikanter Differenzen, im Bereich von 0.14 bis 0.21 über den gesamten Zeitraum betrachtet. Zwischen den Positionen 2 und 3 steigt diese Spannweite leicht an (0.23 bis 0.39 ohne Soziologie innerhalb der Perioden durchgängig signifikanter Differenzen (0.24 bis 0.32 im gesamten Zeitraum). Die Übergänge von Position 3 zu 4 (0.34 bis 0.40/0.24 bis 0.30) und von 4 zu 5 (0.33 bis 0.50 ohne Soziologie/0.37 bis 0.57) weisen ein ähnliches Steigungsverhältnis auf. Zwischen den Positionen 5 und 6 beträgt die mittlere Distanz im Notenniveau schließlich 0.26 (0.22) Noten, zwischen 6 und 7 0.58 (0.58) Noten.

Innerhalb der Positionen, auf denen sich mehrere Studiengänge befinden reicht die Spannweite an durchschnittlichen Distanzen, hier jeweils als Betragsfunktion der jährlichen Differenzen über den gesamten Zeitraum gemittelt, von 0.07 Noten (zwischen Mathematik und Chemie) bis 0.24 Noten (zwischen Deutsch Lehramt und Soziologie). Zwischen diesen Werten liegen Biologie und Psychologie (0.10), Germanistik und Maschinenbau (0.09)/ Germanistik und Soziologie (0.15)/ Maschinenbau und Soziologie (0.14) sowie Deutsch Lehramt und Mathematik Lehramt (0.13)/Mathematik Lehramt und Soziologie (0.22).

Über den gesamten Zeitraum betrachtet lässt sich demnach ein durchschnittlicher Abstand bis 0.24 Noten als Spielraum innerhalb einer gemeinsamen Position betrachten. Ab 0.14 Noten Abstand im Mittel ist bereits eine Abgrenzung der Positionen möglich, wodurch sich ein Überschneidungsbereich von 0.10 Noten ergibt. Durch den Vergleich der Abstände im Notenniveau zwischen den Positionen, die sich innerhalb der Perioden durchgängig signifikanter Differenzen ergeben mit den Abständen innerhalb der Positionen, lässt sich die Grenze zwischen gemeinsamem und unterschiedlichem No-

[58] Es wird an dieser Stelle die Betragsunktion der Differenzen verwendet, da die Abstände zwischen den Noten von Interesse sind, welche der Mittelwert der Differenzen bei wechselnder Rangfolge im Notenniveau vor allem bei der Berechnung der Abstände innerhalb der Positionen nicht widerspiegelt.

tenniveau enger eingrenzen. Der Schwellenwert, bei dem im hier vorgenommenen Bestimmungsverfahren die Grenze zwischen gemeinsamem und signifikant unterschiedlichem Notenniveau liegt, kann dann im Bereich 0.23 bis 0.24 Noten Abstand identifiziert werden. Dieser empirische Wert liegt damit etwas über dem zuvor angegebenen Beispielwert von 0.14 als Grenze zur statistischen Signifikanz auf 5% Niveau zwischen dem Notenniveau zweier Studiengänge unter den modellhaften Bedingungen einer Fallzahl von n=50 und einer Standardabweichung von σ=0.5.

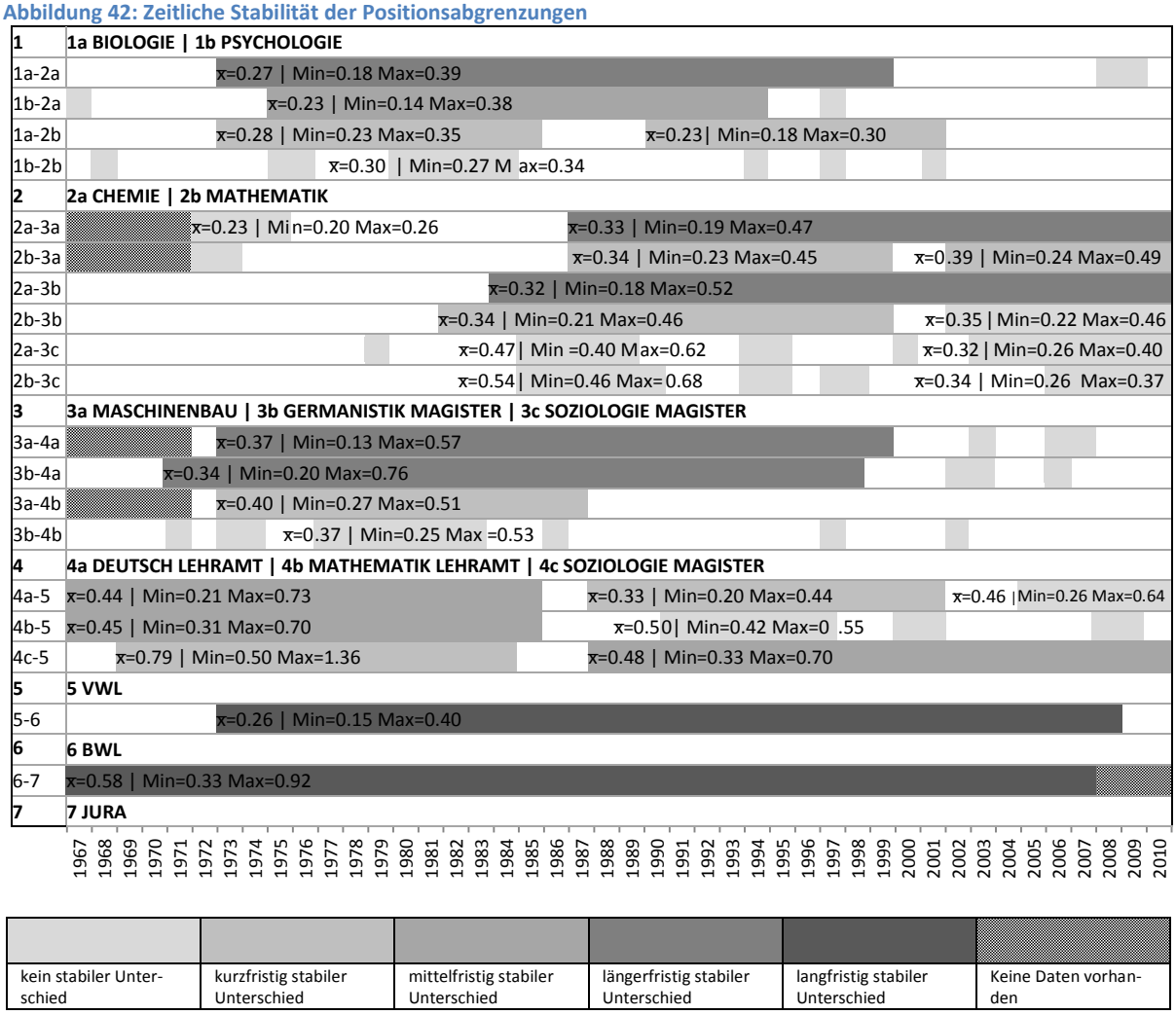
Tabelle 14: Anzahl Jahre mit signifikant differentem Notenniveau zwischen den Studiengängen und Dauer der signifikanten Differenz (Paarvergleiche)

	Biologie	Psychologie	Chemie	Mathematik	Maschinenbau	GermanistikMA	SoziologieMA	MathematikLA	DeutschLA	VWL	BWL	Jura
Jura	1967-2007 40 | 41	1967-2007 40 | 41	1967-2007 40 | 41	1967-2007 40 | 41	1972-2007 35 | 36	1967-2007 38 | 39	1969-2007 38 | 39	1967-2007 40 | 41	1967-2007 40 | 41	1967-2007 40 | 41	1967-2007 40 | 41	--
BWL	1967-2010 43 | 44	1967-2010 43 | 44	1967-2010 43 | 44	1967-2010 43 | 44	1972-2010 38 | 39	1969-2010 41 | 42	1969-2010 41 | 42	1967-2004° 37 | 38	1967-2010 43 | 44	1973-2008° 35 | 34	--	1967-2007 40 | 41
VWL	1967-2010 43 | 44	1967-2010 43 | 44	1967-2010 43 | 44	1967-2010 43 | 44	1972-2010 38 | 39	1969-2010 41 | 42	1969-1984 1988-2010° 15; 22 | 38	1967-1985 1991-1997 18; 6 | 32	1967-1985 1988-2001° 2005-2010 18; 13; 5 | 38	--	1973-2008° 35 | 34	1967-2007 40 | 41
DeutschLA	1969-2010 41 | 42	1969-2009 40 | 41	1967-2005 38 | 41	1967-1999 2002-2009° 32; 7 | 40	1973-1999° 26 | 29	1971-1998°°° 27 | 28	xx 0 | 5	xx 0 | 0	--	1967-1985 1988-2001° 2005-2010 18; 13; 5 | 38	1967-2010 43 | 44	1967-2007 40 | 41
MathematikLA	1969-1999 30 | 35	1968-1999 31 | 34	1967-1989 1994-1999 22; 5 | 31	1967-1989 1993-1999 22; 6 | 32	1973-1987 14 | 15	1977-1983° 6 | 12	xx 0 | 2	--	xx 0 | 0	1967-1985 1991-1997 18; 8 | 32	1967-2004° 37 | 38	1967-2007 40 | 41
SoziologieMA	1973-2000°° 2003-2010° 27; 7 | 33	1975-2000°° 2003-2010°° 25; 7 | 29	1985-1989 2003-2010°° 4;7 | 15	1985-1990 2006-2010° 5; 4 | 14	xx 0 | 4	xx 0 | 1	--	xx 0 | 2	xx 0 | 5	1969-1984 1988-2010° 15; 22 | 38	1969-2010 41 | 42	1967-2007 40 | 41
GermanistikMA	1978-2010 32 | 37	1975-2010°° 35 | 34	1984-2010°° 26 | 25	1982-1999°° 2002-2010 17; 8 | 25	xx 0 | 6	--	xx 0 | 1	1977-1983° 6 | 12	1971-1998°°° 27 | 28	1969-2010 41 | 42	1969-2010 41 | 42	1967-2007 40 | 41
Maschinenbau	1972-2010 38 | 39	1972-2010 38 | 39	1972-1975 1987-2010 3; 23 | 28	1987-1999 2002-2010 12; 8 | 24	--	xx 0 | 6	xx 0 | 4	1973-1987 15 | 16	1973-1999° 26 | 29	1972-2010 38 | 39	1972-2010 38 | 39	1972-2007 35 | 36
Mathematik	1973-1985 1990-2001 12; 11 | 25	1980-1984 4 | 11	xx 0 | 1	--	1987-1999 2002-2010 12; 8 | 24	1982-1999°° 2002-2010 17; 8 | 25	1985-1990 2006-2010° 5; 4 | 14	1967-1989 1993-1999 22; 6 | 32	1967-1999 2002-2009° 32; 7 | 40	1967-2010 43 | 44	1967-2010 43 | 44	1967-2007 40 | 41
Chemie	1973-1999°° 26 | 27	1975-1994°°° 19 | 19	--	xx 0 | 1	1972-1975 1987-2010 3; 23 | 28	1984-2010°° 26 | 25	1985-1989 2003-2010°° 4;7 | 15	1967-1989 1994-1999 22; 5 | 31	1967-2005 38 | 41	1967-2010 43 | 44	1967-2010 43 | 44	1967-2007 40 | 41
Psychologie	1990-1998°° 8 | 8	--	1975-1994°°° 19 | 19	1980-1984 4 | 11	1972-2010 38 | 39	1975-2010°° 35 | 34	1977-2000°° 2003-2010°° 25; 7 | 29	1968-1999 31 | 34	1969-2009 40 | 41	1967-2010 43 | 44	1967-2010 43 | 44	1967-2007 40 | 41
Biologie	--	1990-1998°° 8 | 8	1973-1999°° 26 | 27	1973-1985 1990-2001 12; 11 | 25	1972-2010 38 | 39	1978-2010 32 | 37	1973-2000°° 2003-2010° 27; 7 | 33	1969-1999 30 | 35	1969-2010 41 | 42	1967-2010 43 | 44	1967-2010 43 | 44	1967-2007 40 | 41

Jedes ° steht für eine einjährige Lücke (=ein nicht signifikanter Wert zwischen zwei signifikanten Werten) im jeweiligen Zeitraum

a \| b bzw. a; a_i \| b ist zu lesen als: Dauer des Zeitraums (a); der Zeiträume (a; a_i) mit durchgängig signifikantem Unterschied im Notenniveau\| Gesamtzahl signifikant differenter Jahreswerte

Abbildung 42: Zeitliche Stabilität der Positionsabgrenzungen



kein stabiler Unterschied	kurzfristig stabiler Unterschied	mittelfristig stabiler Unterschied	längerfristig stabiler Unterschied	langfristig stabiler Unterschied	Keine Daten vorhanden

Lesehilfe: Die Balken markieren die unterschiedlich lang andauernden Zeiträume mit durchgängig signifikantem Unterschied im Notenniveau zwischen den einzelnen Studiengängen der jeweils aneinandergrenzenden Positionen. Der x̄-Wert gibt die gemittelte Betragsfunktion der jährlichen Differenzen (=Distanz) zwischen diesen Studiengängen wieder, Min und Max die Minimal- und Maximalwerte der Distanz.

Werden die jeweils fünfstufigen Klassierungen für alle Paarvergleiche in Relation zueinander gesetzt, ergeben sich 25 theoretische Kombinationsmöglichkeiten. Wird berücksichtigt, dass die Klasse der Dauer nicht größer sein kann als die Gesamtzahl der Werte[59], bleiben 15 Kombinationsmöglichkeiten übrig: Fünf, in denen die beiden Klasseneinteilungen gleich sind und 10, in denen die Dauer im Vergleich zur Gesamtzahl niedriger eingestuft wird. Aus einer entsprechenden Kreuztabellierung[60] lassen sich drei hervorstechende Gruppen von Paarvergleichen erkennen: 1. Paarvergleiche mit einer hohen Beziehungsklasse (Klasse 5) hinsichtlich der Gesamtzahl signifikant differenter Werte und einer langfristigen Stabilität dieser Differenzen (Klasse 5) 2. Paarvergleiche mit einer niedrigen Beziehungsklasse (Klassen 1 und 2) und einer kurzfristigen oder gar nicht vorhandenen Stabilität der Differenzen

[59] Im Streudiagramm ist zu erkennen, dass in Einzelfällen rechnerisch minimal höhere Werte für die Periodendauer als für die Gesamtzahl erzielt werden, was durch die Berücksichtigung einzelner nicht signifikanter Jahre zwischen zwei signifikanten Jahren in der Berechnung der Periodendauern zustande kommt.

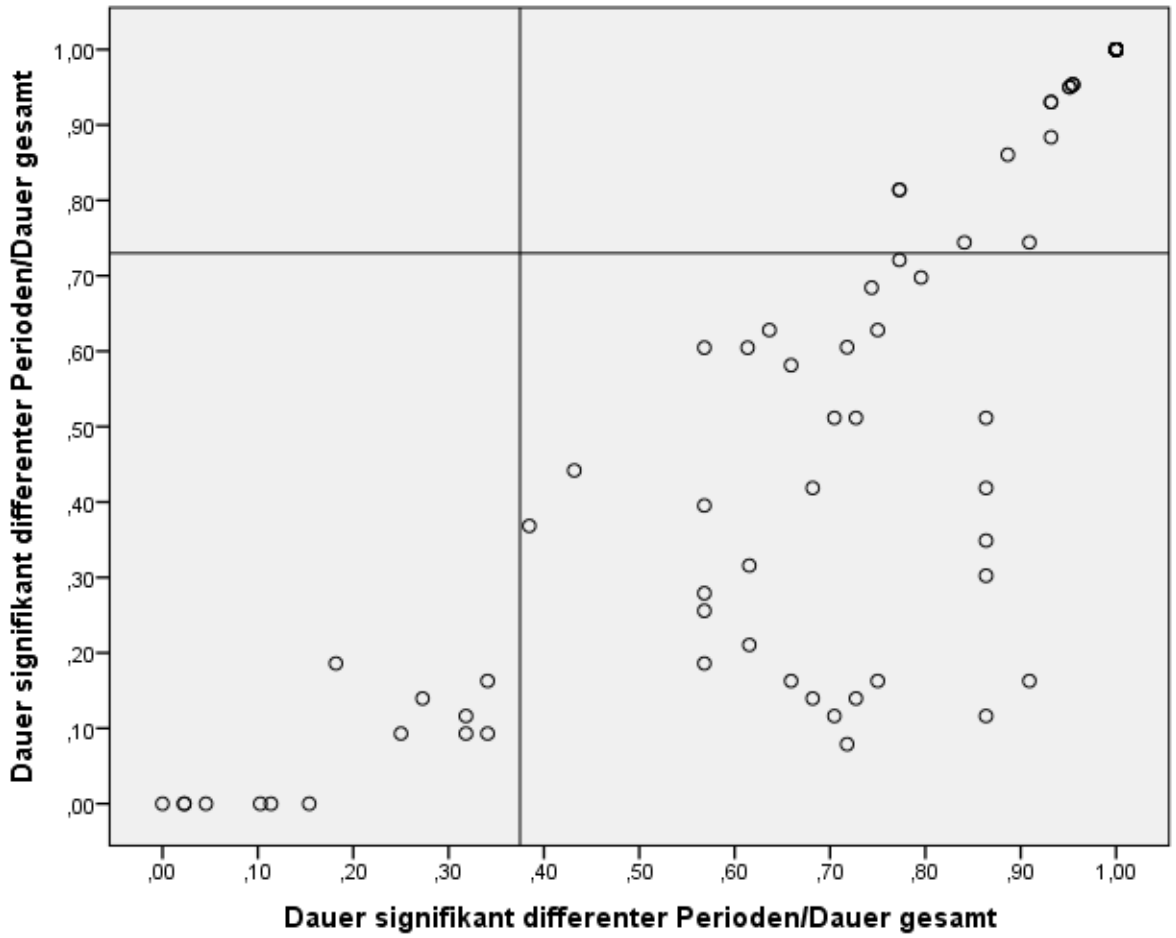
[60] Kreuztabelle (Tab.A16) im Anhang

(Klasse 1) 3. Paarvergleiche mit einer relativ hohen Gesamtzahl an signifikant differenten Werten (Klasse 4), die die gesamte Bandbreite an Periodendauern abdecken (Klassen 1 bis 4).

Die Gegenüberstellung der relationalen Werte der Häufigkeiten signifikant differenter Werte und der Periodendauern in einem Streudiagramm legt diese Unterscheidung ebenfalls nahe. Es zeigt sich wiederum, dass viele signifikante Werte nicht unbedingt auch eine langfristig stabile Differenz zwischen den Studiengängen bedeuten. Neben den beiden naheliegenden Typen von Paarvergleichen mit einer hohen Gesamtzahl signifikant differenter Werte und einer langfristigen Stabilität dieser Differenz sowie einer niedrigen Gesamtzahl signifikant differenter Werte und einer kurzfristigen oder gar nicht vorhandenen Stabilität dieser Differenz lassen sich einige Vergleiche ausmachen, die eine relativ hohe Anzahl an Gesamtwerten und eine mittel- bis längerfristige Beziehung aufweisen, sowie eine Vielzahl Fälle, die gegenüber der relativ hohen Gesamtzahl an signifikant differenten Werten eine vergleichsweise niedrige Stabilität dieser Differenzen aufweisen.

Eine hierarchische Clusteranalyse bestätigt diese Eindrücke. Die Analyse ergibt eine optimale Lösung von drei Clustern. Cluster 1 besteht aus den Paarvergleichen, die neben einer hohen Gesamtzahl auch langfristige Stabilität aufweisen, Cluster 3 umfasst diejenigen Paarvergleiche, welche neben einer niedrigen Gesamtzahl niedrige oder keine Stabilität aufweisen und Cluster 2 enthält sowohl die Paarvergleiche mittelhoher Gesamtwerte und mittlerer Stabilität als auch die Fälle mit mittelhohen und niedrigen Gesamtwerten und vergleichsweise niedrigeren Periodendauern (siehe Anhang: Tab.A17).

Abbildung 43: Streudiagramm Paarvergleiche – Verhältnis Anzahl signifikant differenter Werte zu Periodendauer



Die Querschnittsanalyse über den gesamten Zeitraum seit 1960 hinweg führt zusammengefasst zu folgenden Erkenntnissen: Seit spätestens Anfang der 1970er Jahre herrscht eine stabile Notenhierarchie an deutschen Hochschulen. Die Rangfolge beginnt mit Biologie als Studiengang mit den besten Noten, dicht gefolgt von Psychologie. In den ersten juristischen Staatsexamen werden im Durchschnitt die schlechtesten Noten vergeben, sie sind mehr als doppelt so hoch wie in Biologie. In den beiden wirtschaftswissenschaftlichen Studiengängen BWL und VWL sind die schlechtesten Durchschnittsnoten innerhalb der Diplomstudiengänge zu finden. Die Forschungshypothese FH2a (hochschulübergreifend zeitlich stabile Differenzen im Notenniveau zwischen fachlich abgegrenzten Studiengängen) ist damit bestätigt.

Die Studiengänge mit sehr guten und sehr schlechten Notenniveaus weisen eine niedrige Streuung der Noten auf, für die Studiengänge im mittleren Bereich der Notenskala zeigt sich kein Muster. Anhand der Anzahl an Jahren, in denen sich das Notenniveau zwischen den einzelnen Studiengängen signifikant unterscheidet, lassen sich aus der Rangfolge der 12 Studiengänge sieben Positionen bestimmen, innerhalb derer jeweils ein ähnliches Notenniveau besteht. Die Differenzen im Notenniveau zwischen den Positionen müssen jedoch im zeitlichen Kontext betrachtet werden - ihre zeitliche Stabilität variiert deutlich. Die Grenze zwischen gemeinsamem und signifikant unterschiedlichem Notenniveau liegt im Bereich 0.22 bis 0.24 Noten Abstand. Insgesamt lassen sich drei unterschiedli-

che Beziehungsmuster zwischen den einzelnen Studiengängen ausmachen: 1. Hohe Anzahl signifikant differenter Jahre und hohe zeitliche Stabilität dieser Differenzen, 2. Niedrige Anzahl signifikant differenter Jahre und niedrige oder keine Stabilität dieser Differenzen und 3. Niedrige bis mittelhohe Anzahl signifikant differenter Jahre und niedrige bis mittlere Stabilität dieser Differenzen. Auffällig ist, dass die beiden in der Stichprobe enthaltenen Magister- und die beiden Lehramtsstudiengänge jeweils ein ähnlich hohes Notenniveau aufweisen.

Die Stichprobe gibt in der Mehrzahl der Studiengänge das Notenniveau wieder, das auch andere Quellen aufweisen. Der Vergleich der Werte der Stichprobe mit den Werten von Hitpass/Trosien und dem Wissenschaftsrat sowie mit den Bundesdurchschnitten[61] der FDZ Daten seit 1995 (Anhang: Tab.A18-A19; Abb.A1-A4), zeigt nur leichte Abweichungen: So deuten die Datenvergleiche darauf hin, dass die BWL-Noten in der Stichprobe durchgehend leicht überschätzt werden und vermutlich etwas niedriger, etwa auf VWL-Ebene, liegen.

Für die Differenzierung der Studiengänge nach zeitlich stabilen Notenunterschieden könnte dies bedeuten, dass sich die Abstufung womöglich von sieben auf sechs Positionen verringern lässt, BWL und VWL auf Studiengangebene eher ein über die Zeit gemeinsames Notenniveau aufweisen. In den beiden Lehramtsstudiengängen und in Chemie liegt das Notenniveau in der Stichprobe in der Regel unter den Vergleichswerten. Auf die Einordnung in die Positionshierarchie der Studiengänge dürfte der leichte Bias hier allerdings keine Auswirkungen haben.

8.1.2 Die langfristige Entwicklung des Notenniveaus in den untersuchten Studiengängen

Die Querschnittanalyse der Daten zeigt, dass studiengangspezifische Notenniveaus existieren, welche im Zeitverlauf nicht gleichermaßen stabil sind. Wie in Abbildung 40 bereits zu erkennen ist, ist dies unterschiedlichen langfristigen Entwicklungen der Studiengänge geschuldet.

Welche langfristigen Entwicklungstendenzen sind zu erwarten und welche lassen sich in den Daten finden? Eine intertemporale Vergleichbarkeit vom Verhältnis Leistung zu Note besitzt nur dann uneingeschränkte Aussagekraft, wenn, unabhängig von Zeitpunkt und Ort der Prüfung, für die gleiche Leistung auch die gleiche Beurteilung erfolgt[62]. Unter der modellhaften Annahme, dass sowohl die durchschnittliche Leistungsfähigkeit der Studierenden als auch eventuelle externe Einflüsse auf die Notengebung im Zeitverlauf konstant sind, sollten sich die Noten auf einem konstanten Durchschnittsniveau befinden. Werden die vergebenen Noten immer besser, obwohl die Leistungen von Studierenden und das Leistungsgefälle zwischen ihnen konstant bleiben, wird damit das Prinzip der Vergleichbarkeit ausgehebelt. Absolvent*innen mit guten Leistungen bekommen weiterhin gute No-

[61] Ob die Stichprobenwerte mit den FDZ-Notenmitteln über alle Hochschulen je Studiengang oder nur über westdeutsche Hochschulen je Studiengang verglichen werden, macht keinen nennenswerten Unterschied.

[62] Zudem beinhalten intertemporale Vergleiche bei vorwiegender Nutzung der absoluten Bezugsnorm unterschiedliche Wissensbestände, weshalb intertemporale Vergleiche streng genommen die relative Leistung zum jeweils vorhandenen Wissensstand des Fachs betrachten müssten (vgl. Müller-Benedict/Gaens 2015).

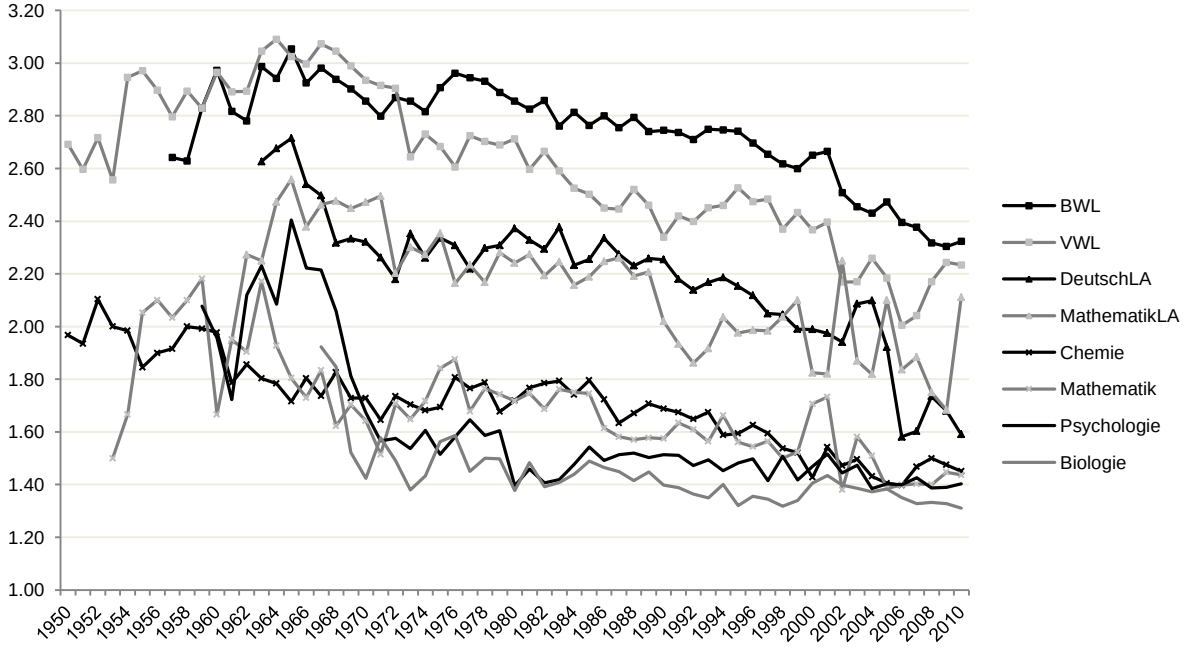
ten, solche mit schlechteren Leistungen dann aber ebenfalls, womit eine Differenzierung zwischen ihnen erschwert wird. Ein sinkender Notenverlauf gilt als Indiz für eine solche grade inflation, die möglicherweise auch zur Problematik von grade compression führt (siehe Abschnitt 6.3.1).

Neben den idealtypisch linearen Verlaufsformen eines konstanten Notenniveaus und einer kontinuierlichen Verbesserung (oder prinzipiell auch Verschlechterung) besteht noch eine Möglichkeit der langfristigen Notenentwicklung: Zyklische Bewegungen. Müller-Benedict und Tsarouha (2011) vermuten, ausgehend vom Befund zyklischer Mangel- und Überfüllungsperioden im Akademiker*innenarbeitsmarkt (Titze 1990) zyklische Auswirkungen auf das Notenniveau (auch: Müller-Benedict 2005 zum Zusammenhang zwischen Akademiker*innenzyklen und Prüfungserfolg).

Im deutschen Kontext sind Müller-Benedict und Tsarouha die ersten, die explizit auf die Möglichkeit zyklischer Notenverläufe hinweisen. Verbesserungsperioden stellen dieser Überlegung nach nicht zwangsläufig einen kontinuierlichen Trend zu besseren Noten dar, sondern sind möglicherweise Bestandteil einer umfassenderen Dynamik. Auch in der US-amerikanischen Forschung, die sich seit den 1970er Jahren mit der langfristigen Entwicklung von Noten im tertiären Bildungsbereich auseinandersetzt, wird die Möglichkeit einer zyklischen Prozessdynamik, wie bereits erwähnt, nur von Kolevzon (1981) überhaupt in Betracht gezogen. Sie stellt damit eine bislang unerforschte Alternativhypothese zur Annahme ausschließlich trenddominierter Verläufe von Noten dar.

Sollten die zahlreichen Warnungen vor einem allumfassenden Notenverfall berechtigt sein, müssten die Zeitreihen auf Studiengangebene vor allem Verbesserungen abbilden. Und tatsächlich zeigen die Daten in acht von 12 berücksichtigten Studiengängen eine nennenswerte Verbesserung der Notendurchschnitte im Zeitverlauf:

Abbildung 44: Zeitlicher Verlauf der Abschlussnoten in Studiengängen mit langfristiger Notenverbesserung



Die Grafik zeigt, dass die Verbesserung jeweils zu Beginn/Mitte der 1960er Jahre, in Mathematik und Chemie etwas früher, einsetzt, was den Daten von Hitpass und Trosien entspricht. Die Verbesserungsprozesse vollziehen sich allerdings auf unterschiedlich hohem Niveau. Die durchschnittliche Abschlussnote in BWL ist 2010 trotz langfristiger Verbesserung immer noch signifikant schlechter als das Notenniveau in Chemie 1960 (p=0.000) und in Biologie 1967 (p=0.009), also bevor dort die Verbesserung auftritt. Die durchschnittliche Abschlussnote in BWL ist 2010 trotz langfristiger Verbesserung immer noch signifikant schlechter als das Notenniveau in Chemie 1960 (p=0.000) und in Biologie 1967 (p=0.009), also bevor dort die Verbesserung auftritt. Die statistische Signifikanz ist in beiden Fällen bedeutsam, der Effekt in beiden Vergleichen als mittelstark einzustufen (Hedges' g=0.548 für Chemie vs. BWL bzw. 0.733 für Biologie vs. BWL). Die Abstufung der Studiengänge nach Notenniveau entspricht für die fünf der sechs Diplomstudiengänge, für die zum Vergleich sowohl Daten von Hitpass und Trosien als auch vom Wissenschaftsrat vorliegen, exakt der dortigen Abstufung.
In Biologie und Psychologie kann man die Notenlage spätestens seit Beginn der 1970er Jahre als derart gut einstufen, dass die Leistungsdifferenzierung dort zwangsläufig durch eine Häufung der Noten im Bestbereich gefährdet ist. Dafür sprechen sowohl die in Richtung der besten Noten verzerrten Notenverteilungen als auch die entsprechend geringen Standardabweichungen der Individualnoten in diesen beiden Studiengängen. In Psychologie wurden 54.6% der 11 467 zwischen 1971 und 1997 bestandenen Prüfungen mit einem „sehr gut“, 95.5% mit einem „sehr gut“ oder „gut“ bewertet (Schiefe=0.713). In Biologie liegen diese Anteile im gleichen Zeitraum bei 62.8% bzw. 96.5% (n=11 611, Schiefe=1.033). Der Wert der Standardabweichungen beträgt in Psychologie seit 1971 in jedem Jahr zwischen s=0.53 und s=0.61, in Biologie seit 1977 zwischen s=0.49 und s=0.63. Nur die Noten in Maschinenbau streuen ähnlich gering (zwischen s=0.45 und s=0.62 seit 1973), was in diesem Fall aufgrund des höheren Notenniveaus aber nicht durch die Begrenzung der Notenskala erklärt werden kann. Auch das Ausmaß der Verbesserung variiert. Während sich der Durchschnitt in Chemie in 50 Jahren ca. eine halbe Note senkt, fällt der Durchschnitt in Deutsch mit Abschluss Lehramt um mehr als eine ganze Note und damit um etwa das Doppelte. Auch in VWL und Psychologie sind die Noten in den 2000er Jahren im Mittel mehr als eine ganze Note besser als zu Beginn des Abwärtstrends. Es folgen Mathematik Lehramt und Diplom, dann BWL und Biologie (Tab.15).

Gemeinsam ist allen Studiengängen, dass die Verbesserung nicht komplett durchgängig verläuft. Sie findet in Phasen von unterschiedlicher Länge statt, die von Plateauphasen unterbrochen werden. In Biologie ist der Großteil der Verbesserung, die sich 2010 im Vergleich zu 1967 feststellen lässt, bereits 1973 erreicht. Sechs Jahre nachdem die verfügbare Zeitreihe beginnt, ist das Notenniveau von $\bar{x}$=1.92 auf $\bar{x}$=1.38 gesunken. Es lässt sich anhand der vorliegenden Daten nicht genau sagen, wann dieser Prozess in Biologie eingesetzt hat und wie lange er anhielt. Da der Diplomabschluss in Biologie

an den meisten Hochschulen aber erst in den 1960er Jahren eingeführt wurde, lässt sich auch ohne weiter zurückgehende Daten folgern, dass sich der Verbesserungsprozess in wenigen Jahren vollzogen haben muss. In den anderen Studiengängen lassen sich mehrere maßgebliche Verbesserungsperioden ausmachen: In Chemie sinkt das Notenniveau zwischen 1958 und 1971 (-0.35) sowie zwischen 1987 und 2006 (-0.23). Die maximale in der Reihe enthaltene Verbesserungsspanne von -0.70, die über eine Dauer von 54 Jahren (1952-2006) festzustellen ist, beruht also größtenteils auf zwei Phasen der Verbesserung, die in zusammen 32 Jahren für eine Absinken des Notenniveaus um 0.58 Noten verantwortlich sind. In Psychologie findet die größte Verbesserung zwischen 1965 und 1971 statt (-0.83), eine zweite, sehr kurze Verbesserungsperiode lässt sich zwischen 1979 und 1982 beobachten (-0.19). Zwischen 1971 und 1979 wie auch zwischen 1982 und 2010 bewegen sich die Durchschnittsnoten in einem relativ stabilen Rahmen.

Auch die wirtschaftswissenschaftlichen und die Lehramtsstudiengänge weisen das Muster einer phasendominierten Verbesserung auf: In BWL sinkt das Notenniveau zwischen 1965 und 1971 und noch einmal zwischen 1984 und 2009. In VWL sind es mit den Jahren 1967-1973, 1982-1990 und 2001-2006 drei Phasen, die sich hauptsächlich für die langfristige Verbesserung im Studiengang verantwortlich zeichnen. In den beiden Lehramtsstudiengängen verlaufen die Entwicklungen beinahe parallel: In Deutsch verbessern sich die Noten wesentlich zwischen 1965-1970 und zwischen 1986-2006, in Mathematik zwischen 1965-1972 und zwischen 1989-2009, wobei in Mathematik am Ende der Zeitreihe einige Ausreißerwerte das Bild verzerren. Der Verlauf in Mathematik Diplom weicht von der in den anderen Studiengängen zu beobachtenden Verlaufsform der durch (eine) Plateauphase(n) unterbrochenen Verbesserung ab. Zwar lassen sich mit den Zeiträumen von 1963-1971 und von 1985-2002 ebenfalls zwei maßgebliche Verbesserungsperioden bestimmen, jedoch folgt auf die erste dieser Phasen zunächst wieder eine Verschlechterung bis 1976, bevor sich eine neunjährige Plateauphase anschließt.

Tabelle 15 fasst die Verbesserungs- und Plateauphasen zusammen. Spalte 2 gibt die maßgeblichen Verbesserungsperioden an, also die Zeiträume, in denen die Noten den stärksten Abwärtstrends unterliegen, in den Klammern findet sich das Ausmaß der jeweiligen Verbesserung in dieser Zeit. Dass das Ausmaß der Verbesserung als absolute Angabe zwischen den Studiengängen sinnvoll vergleichbar ist, zeigt sich durch einen Abgleich der Werte mit der jeweiligen Effektstärke *g* (die Effektstärken befinden sich im Anhang: Tab.A20): Die Korrelation zwischen den beiden Kennwerten beträgt r=0.964 (p=0.000). In Spalte 3 sind die Plateauphasen und in Klammern die Spannweiten, innerhalb derer sich das Notenniveau in diesen Phasen bewegt, verzeichnet. In Chemie etwa beträgt die Differenz zwischen höchster und niedrigster Durchschnittsnote zwischen 1971 und 1987 0.13 Noten. Die Plateauphasen zeichnet aus, dass sie eine Spannweite von R=0.20 nicht überschreiten.

Die Spalten 4 und 5 bieten einen Vergleich der maximal in der Zeitreihe zu beobachtenden Verbesserung, also der Differenz zwischen höchstem und niedrigstem Notendurchschnitt, und der Zeitspanne in der dieses Maximum erreicht wird (Spalte 4) mit dem Ausmaß, das nur in den beschriebenen Verbesserungsperioden (Spalte 5) zustande kommt[63]. Hier zeigt sich, dass die größte Verbesserung über den gesamten Zeitraum in Deutsch (Lehramt), die geringste in Biologie stattgefunden hat (Spalte 4). Die letzte Spalte enthält zwei Werte, die die durchschnittliche Verbesserung im Notenniveau pro Jahr abbilden: Der erste Wert über den gesamten Zeitraum seit Einsetzen der ersten Verbesserung bis zum letzten Messzeitpunkt 2010, der zweite Wert nur für die Jahre, die in die Verbesserungsphase(n) fallen (siehe Spalte1). Berechnet sind diese Angaben als arithmetisches Mittel der 1.Differenzen der entsprechenden Datenpunkte, also als Durchschnittswert der jährlichen Veränderungen gegenüber den Vorjahreswerten [64].Der über den gesamten Zeitraum gemittelte Wert gibt ähnlich wie der absolute Wert in Spalte 4 das Ausmaß dieser Verbesserung seit dem ersten Einsetzen der Verbesserung an, nur eben nicht begrenzt auf die Zeitspanne der maximalen Verbesserung. Am zweiten Wert, dem Durchschnitt innerhalb der Verbesserungsphase(n), lässt sich ablesen, wie stark die Noten in den Jahren sinken, in denen die Verbesserung maßgeblich stattfindet.

Spalte 6 lässt erkennen, dass sich die Notenveränderung in Psychologie als besonders dynamisch erweist: Dort verbessert sich das Niveau während der Verbesserungsphasen mit Abstand am stärksten. Es wird außerdem auch hier wieder deutlich, dass die einfache Differenz eines Anfangs- und Endwerts die Dynamik zwischen diesen Messpunkten in der Regel verschweigt: Die jährlichen Verbesserungen in diesen Phasen liegen in allen Fällen über den Durchschnittswerten für den gesamten Zeitraum. Der Vergleich des Zeitraums der maximalen Verbesserung (Spalte 4) mit der summierten Dauer der Verbesserungsphasen (Spalte5) veranschaulicht die Dauer, über die tatsächlich dynamische Bewegungen in den Noten zu beobachten sind. Es fällt auf, dass in allen Studiengängen, in denen sich eine langfristige Verbesserung zeigt, die erste dafür maßgeblich verantwortliche Phase bereits Anfang der 1970er abgeschlossen ist. Das folgende Plateau besteht je nach Studiengang zwischen acht und 17 Jahren, die zweite Verbesserungsphase setzt entsprechend versetzt Mitte/Ende der 1980er Jahre ein (in Psychologie bereits 1979).

[63] Größere Werte für letztere als für die maximale Gesamtverbesserung kommen dadurch zustande, dass die zweite Periode der Verbesserung leicht über dem Wert wieder einsetzt, mit dem die erste Periode abschließt. In Mathematik (Diplom) liegt dieser Wert deshalb deutlich höher, weil der Ausgangswert der zweiten Verbesserungsperiode dem Peak des zwischenliegenden Zyklus entspricht, der weit über dem Plateauniveau liegt.

[64] Auf diese Weise ergibt sich ein Durchschnittswert, der eine genauere Einschätzung der durchschnittlichen jährlichen Entwicklung bietet als bspw. eine Regression der (nicht perfekt linear verlaufenden) Noten auf eine Zeitvariable.

Tabelle 15: Verlaufsphasen und Verbesserungsausmaß in den Studiengängen mit langfristiger Notenverbesserung

Studiengang	Phasen der Verbesserung (Ausmaß)	Plateauphasen (Spannweite der Schwankungen)	Maximale Verbesserung (Wert/Jahre)	Verbesserung in den Phasen (Wert/Jahre)	ø Verbesserung pro Jahr (Gesamt/ Verbesserungsphasen)
Biologie Diplom	1967-1973 (-0.54**)	1973-2010 (0.19)[a]	-0.61***/43	-0.54/6	-0.014/-0.091
Psychologie Diplom	1965-1971 (-0.83***) 1979-1982 (-0.19***)[b]	1971-1979 (0.13) 1982-2010 (0.15)	-1.01***/39	-1.02/9	-0.022/-0.115
VWL Diplom	1967-1973 (-0.43***) 1982-1990 (-0.32***) 2001-2006 (-0.40***)	1973-1982 (0.13) 1990-2001 (0.19)	-1.09***/42	-1.15/19	-0.020/-0.060
Mathematik Diplom	1963-1971 (-0.66***) 1985-2002 (-0.37***)	1971-1976 (+0.37)[c] 1976-1985 (0.20)	-0.79***/39	-1.03/25	-0.016/-0.041
Deutsch Lehramt	1965-1970 (-0.39***) 1986-2006 (-0.76***)	1970-1986 (0.20)	-1.13***/41	-1.15/25	-0.025/-0.046
Mathematik Lehramt	1965-1972 (-0.36***) 1989-2009 (-0.53***)	1972-1989 (0.19)	-0.88***/44	-0.89/27	-0.010/-0.033
BWL Diplom	1965-1971 (-0.25***) 1984-2009 (-0.51***)	1971-1984 (0.20)	-0.75***/44	-0.76/31	-0.016/-0.025
Chemie Diplom	1958-1971 (-0.35***) 1987-2006 (-0.23***)	1971-1987 (0.13)	-0.70***/54	-0.58/32	-0.011/-0.019

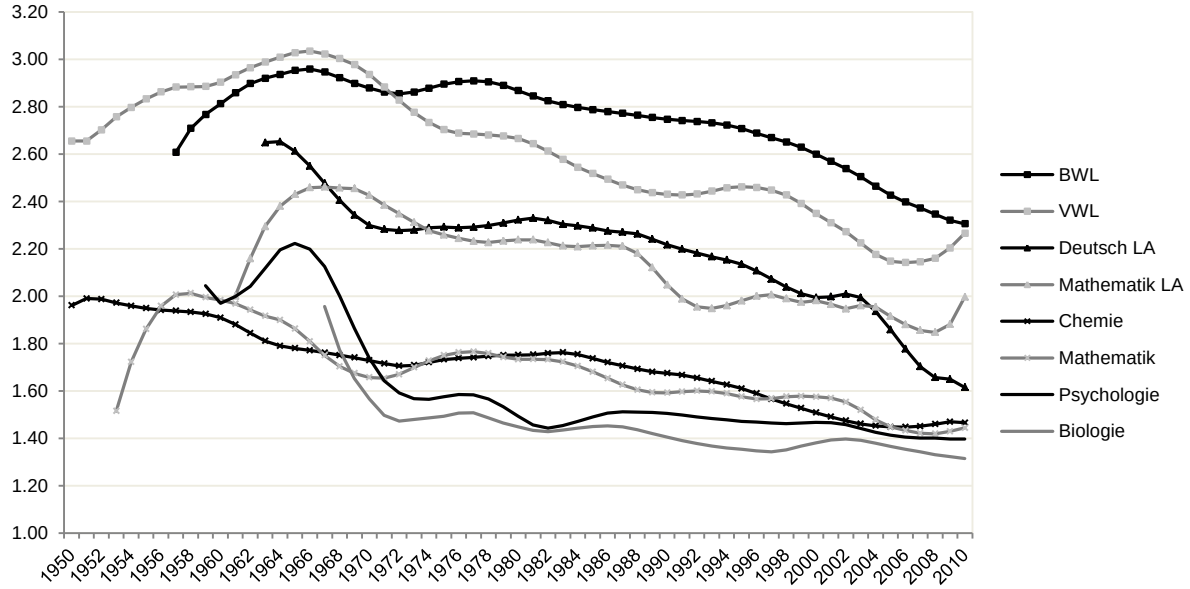
**p≤0.01 ***p≤0.001 [a] ohne Werte für 1975 und 1976 [b] ohne Wert für 1980 [c] Verschlechterungsperiode

Die langfristige Verbesserung der Noten in den acht dargestellten Studiengängen vollzieht sich also phasenweise. Die bisherige Analyse legt nahe, dass sich Phasen der Verbesserung mit Phasen der Konstanz abwechseln, wodurch langfristig eine Verbesserung des Notenniveaus erfolgt. Es gibt allerdings mit Mathematik Diplom einen Studiengang, der nicht ganz in dieses Muster passt. Hier verschlechtert sich das Notenniveau nach der ersten Verbesserungsperiode. Inmitten der langfristigen Verbesserung scheint eine zyklische Komponente zu existieren, die das Absinken des Notenniveaus verzögert. Um zu klären, ob die identifizierten Plateauphasen tatsächlich ein konstantes Notenniveau darstellen oder ob sie nicht doch eher, wie es die Zeitreihe der Mathematiknoten nahelegt, Teil eines zyklischen Verlaufs sind, werden die Zeitreihen nun mit Instrumenten der Zeitreihenanalyse behandelt.

In Abbildung 45 sind die Zeitreihen der Noten in den Studiengängen mit langfristiger Verbesserung noch einmal dargestellt, hier allerdings mit der LOWESS Technik geglättet. LOWESS steht für „**LO**cally **WE**ighted regression **S**catterplot **S**moothing“, ein Anpassungsverfahren für Daten, das die grafische Analyse des Zusammenhangs zweier metrischer Variablen erleichtert. Die Glättung der Daten wird, wie der Name vermuten lässt, durch ein Regressionsmodell erreicht. Es handelt sich dabei um eine polynomiale Regression, für die der Zusammenhang zwischen den Daten nicht vor der Berechnung bekannt sein muss, sondern der erst durch die Berechnung bestimmt wird. Festgelegt wird vor Anwendung dieses nicht-parametrischen Regressionsmodells lediglich die Bandbreite der benachbarten Datenpunkte, die in die Glättung jedes einzelnen Werts eingehen. Je höher sie gewählt wird, umso

glatter werden die Daten (Cleveland 1979; Wolf/Best 2010). Alle im Folgenden als geglättet bezeichneten Daten wurden mit dem Verfahren der LOWESS-Glättung behandelt, indem die Durchschnittsnoten als abhängige, die Zeitvariable als unabhängige Variable in das Regressionsmodell aufgenommen wurden. Die jeweils vorgegebene Bandbreite der Glättung ist stets angegeben.

Abbildung 45: Zeitlicher Verlauf der Abschlussnoten in Studiengängen mit langfristiger Notenverbesserung (LOWESS 0.2)



Neben dem bereits aus den nicht geglätteten Daten erkennbaren Abwärtstrend in den Reihen verdeutlicht die Glättung zuvor nicht erkennbare Schwankungen im Zeitverlauf. Nicht nur in Mathematik Diplom, auch in BWL, VWL, Mathematik Lehramt, Psychologie und Biologie lassen sich jetzt zyklische Bewegungen erkennen. Um diese Zyklen in den Fokus nehmen zu können, werden die geglätteten Reihen in einem nächsten Schritt trendbereinigt. Die Trendbereinigung erfolgt in zwei Schritten. Zunächst wird mit Hilfe der LOWESS-Technik die Trendkomponente bestimmt (Abb.46). Im Gegensatz zur Trendbestimmung über eine lineare Regression der Noten auf die Zeitvariable wird mithilfe der LOWESS Technik nicht nur der (bei der OLS unterstellte) lineare Trend verdeutlicht, sondern ein dynamischer Trendverlauf aufgezeigt, der sowohl lokale Abwärts- als auch Aufwärtstrends sichtbar werden lässt. Die auf diese Weise bestimmte Trendkomponente wird dann von der geglätteten Reihe subtrahiert, wodurch nur noch die zyklische Restkomponente[65] übrig bleibt. Bei nicht perfekt linearen Trends erzielt dieses Verfahren der Trendbereinigung ein genaueres Abbild der tatsächlichen zyklischen Schwankungen um den Trend als beispielsweise die Berechnung der Residuen einer OLS-Regression der Noten auf die Zeitvariable.

[65] Die durch Trendbereinigung erhaltene Reihe umfasst neben der zyklischen Komponente der Original-Zeitreihe noch die sogenannte Restkomponente, die nicht systematische Einflüsse (Störungen) darstellt.

Abbildung 46: Trendkomponenten der Durchschnittsnoten in den acht Studiengängen mit Verbesserung (LOWESS 0.9)

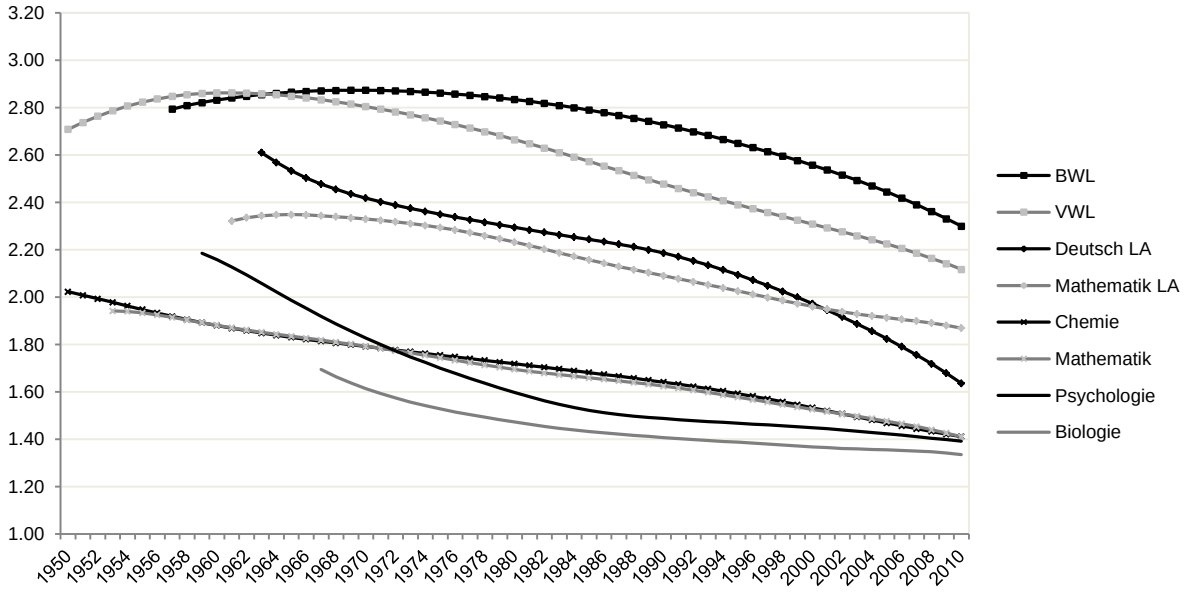


Tabelle 16: Trendstärken der einzelnen Zeitreihen im Vergleich (gemittelte 1. Differenzen der Trendkomponente)

Studiengang	Gesamte Zeitreihe	Nur Abwärtstrend	Bis 1980	Ab 1981	1951-1960	1961-1970	1971-1980	1981-1990	1991-2000	2001-2010
Deutsch Lehramt	-0.021	-0.021 (1963-2010	-0.019	-0.022	--	-0.027 (1964-1970)	-0.012	-0.011	-0.021	-0.034
Psychologie Diplom	-0.016	-0.013 (1966-2010)	-0.028	-0.007	--	-0.033	-0.023	-0.011	-0.004	-0.006
VWL Diplom	-0.010	-0.016 (1962-2010)	-0.001	-0.018	+0.015	-0.006	-0.014	-0.019	-0.017	-0.019
Chemie Diplom	-0.010	-0.10 (1950-2010)	-0.010	-0.010	-0.014	-0.009	-0.007	-0.008	-0.011	-0.012
BWL Diplom	-0.009	-0.015 (1970-2010)	+0.002	-0.018	+0.013 (1958-1960)	+0.004	-0.004	-0.011	-0.017	-0.026
Mathematik Lehramt	-0.009	-0.011 (1966-2010)	-0.005	-0.012	--	+0.001 (1962-1970)	-0.010	-0.014	-0.013	-0.009
Mathematik Diplom	-0.009	-0.010 (1955-2010)	-0.009	-0.009	-0.009 (1954-1960)	-0.009	-0.010	-0.007	-0.010	-0.012
Biologie Diplom	-0.008	-0.008 (1967-2010)	-0.017	-0.005	--	-0.027 (1968-1970)	-0.014	-0.007	-0.004	-0.003

Die Trendbestimmung bestätigt, was bereits die Originaldaten haben erkennen lassen. Alle acht hier berücksichtigten Studiengänge weisen einen klaren Abwärtstrend auf, der sich in unterschiedlicher Stärke und Kontinuität darstellt. Wird die Trendstärke für die einzelnen Studiengänge als arithmetisches Mittel der ersten Differenzen der Trendkomponente berechnet (Tab.16), zeigt sich, dass über die gesamte Reihe Deutsch Lehramt und Psychologie den stärksten Trend aufweisen - in Deutsch Lehramt ist die Abwärtsbewegung im Durchschnitt doppelt so stark, in Psychologie anderthalb mal so stark wie in den übrigen Studiengängen, die gemittelt etwa die gleiche Trendstärke aufweisen (Spalte 2) - allerdings über einen unterschiedlich langen Zeitraum. Dass dies eine Rolle spielt, begründet sich dadurch, dass der Abwärtstrend etwa in BWL und VWL erst Anfang bzw. Mitte der 1960er Jahre einsetzt, während vorher noch ein Aufwärtstrend herrscht.

Da für gut die Hälfte der Studiengänge für den Zeitraum vor 1960 keine Daten vorliegen, ist für sie im Mittelwert über die gesamte Reihe eine solche Verschlechterungsperiode nicht enthalten. Für den Vergleich der Stärke des reinen Abwärtstrends ist es daher sinnvoller, für die Studiengänge mit ei-

nem dem Abwärtstrend vorgelagerten Anstieg der Noten erst ab dem Zeitpunkt zu mitteln, ab dem der Trend sich umkehrt (Spalte 3). Nun zeigt sich, dass der ‚reine' Abwärtstrend in BWL und VWL in etwa so stark ist wie in Psychologie. Mit Deutsch Lehramt als trendstärkstem Studiengang, den beiden wirtschaftswissenschaftlichen Studiengängen und Psychologie lässt sich damit für die Hälfte der acht Zeitreihen mit langfristiger Verbesserung eine deutlich stärkere Trendbelastung feststellen als für die andere Hälfte, die aus Mathematik (Diplom und Lehramt), Chemie und Biologie besteht.

Neben einer unterschiedlichen Gesamtstärke lassen sich außerdem Unterschiede in der zeitlichen Dynamik der Trendstärken ausmachen. Während die Abwärtsbewegung in Mathematik und Chemie beinahe perfekt linear verläuft, weisen BWL und VWL auf der einen und Psychologie und Biologie auf der anderen Seite eine gegenläufige Entwicklung auf: Erstere weisen einen zunächst schwachen, dann immer stärker werdenden, Letztere einen erst sehr starken, dann abschwächenden Abwärtstrend auf. Als Wendepunkt ist der Zeitraum um ca. 1980 zu erkennen. Wird die Trendstärke getrennt für die Zeiträume bis 1980 und danach berechnet (Spalten 4 und 5), bestätigt sich dieser visuelle Eindruck. Nicht nur für Mathematik und Chemie, auch für Deutsch Lehramt zeigt sich kein Unterschied zwischen den beiden Zeiträumen, während Mathematik Lehramt wie BWL und VWL nach 1980 den stärkeren Abwärtstrend aufweist.

Eine Aufteilung in 10-Jahres-Abschnitte (Spalten 6 bis 11) gibt schließlich ein noch genaueres Bild über den zeitlichen Verlauf der Trends. Während in BWL über die Dekaden hinweg die Trendstärke zunimmt, zeigt sich in VWL nach vorheriger Zunahme in den 1990ern ein leichtes Abschwächen, gefolgt von einer erneuten Zunahme in den 2000ern. Mathematik Diplom weist einen ähnlichen Verlauf auf, hier ist das Minimum der Trendstärke allerdings schon in den 1980ern erreicht, die Veränderungen im Ausmaß sind zudem viel geringer als in VWL. In Deutsch Lehramt zeigt sich ein Kurvenverlauf über die 10-Jahres-Werte, der mit einem starken Trend in den 1960ern beginnt, bis in die 1990er ab- und anschließend wieder zunimmt. Alle vier erreichen in den 2000ern die stärkste Phase des Trends. In Mathematik Lehramt ist es genau andersherum, hier nimmt die Trendstärke zunächst bis in die 1980er, in denen sie ihren Höhepunkt erreicht, zu und danach wieder ab. Chemie weist einen ähnlichen Verlauf auf wie Deutsch Lehramt, allerdings mit wesentlich geringeren Veränderungen und der stärksten Phase schon in den 1950ern. In Psychologie und Biologie nimmt die Trendstärke erwartungsgemäß stark ab, sie ist in den 1960ern am stärksten, verliert bis in die 1990er an Kraft und nimmt in Psychologie nur noch einmal leicht zu, während sie in Biologie stagniert.

Die Darstellung der zyklischen Komponenten (Abb.47-54) bestätigt, was sich bereits an den geglätteten Daten erkennen ließ. In allen acht Studiengängen existieren mehr oder weniger starke Zyklen von ca. 20 bis 30 Jahren Länge, wie auch Analysen der Spektraldichte der Reihen bestätigen[66].

[66] Die zugehörigen Periodogramme finden sich im Anhang (Abb.A5-A12).

Abbildung 47: Zyklische Komponente Mathematik (LOWESS 0.3 – LOWESS 0.9)

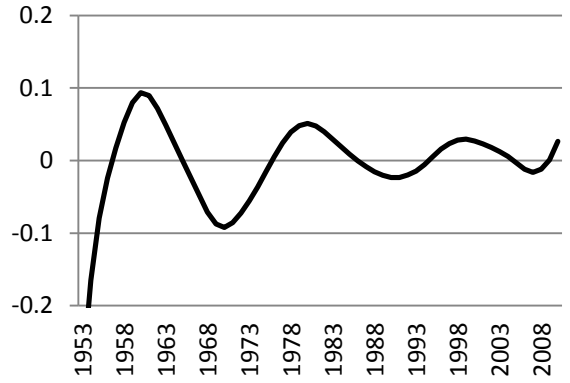


Abbildung 48: Zyklische Komponente Mathematik Lehramt (LOWESS 0.3 – LOWESS 0.9)

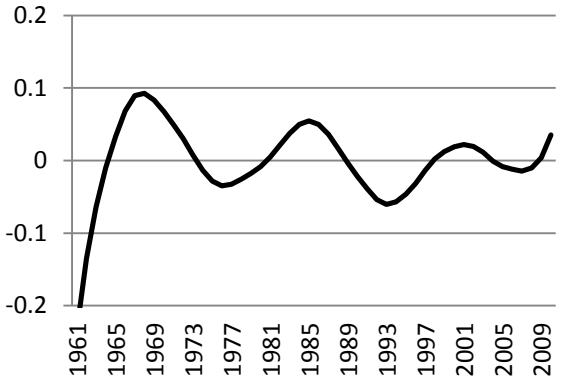


Abbildung 49: Zyklische Komponente Chemie (LOWESS 0.3 – LOWESS 0.9)

Abbildung 50: Zyklische Komponente Biologie (LOWESS 0.4 – LOWESS 0.9)

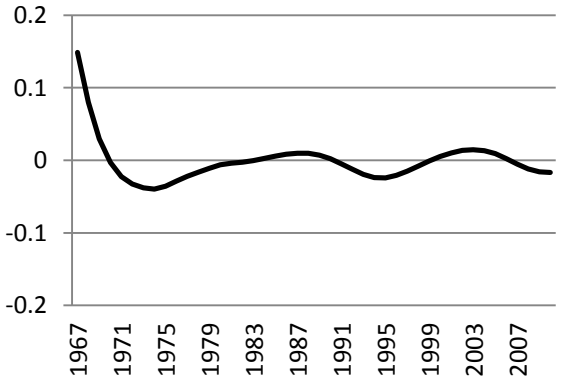


Abbildung 51: Zyklische Komponente VWL (LOWESS 0.4 – LOWESS 0.9)

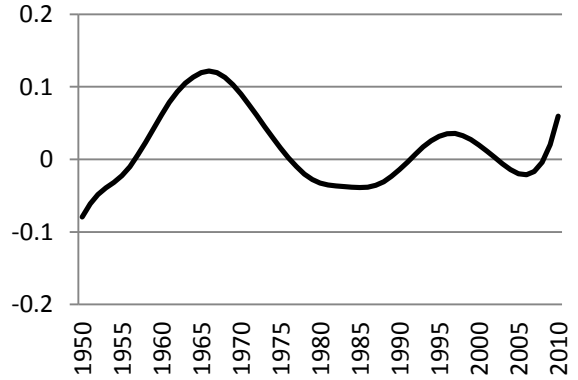


Abbildung 52: Zyklische Komponente BWL (LOWESS 0.3 – LOWESS 0.9)

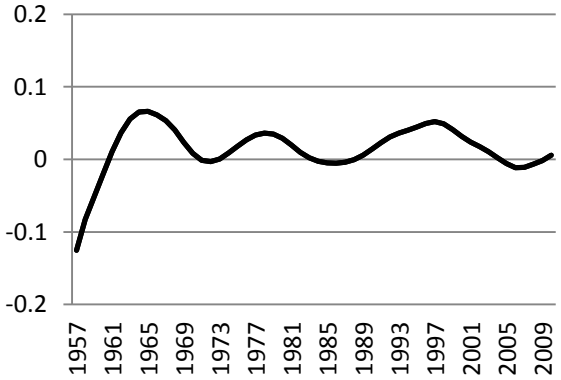


Abbildung 53: Zyklische Komponente Psychologie (LOWESS 0.4 – LOWESS 0.9)

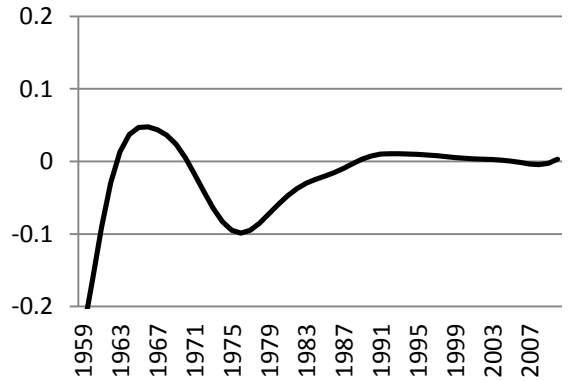
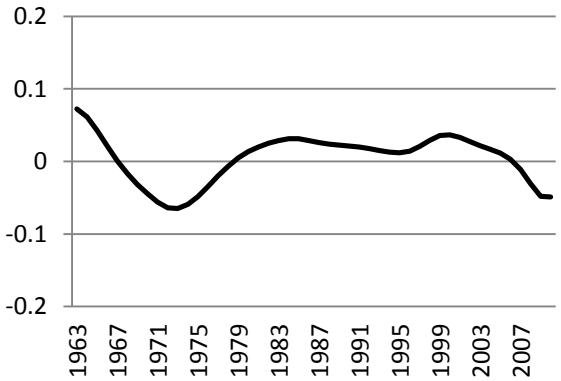


Abbildung 54: Zyklische Komponente Deutsch Lehramt (LOWESS 0.4 – LOWESS 0.9)



Eine weitere Umformung liefert nun schließlich die Antwort auf die Frage, ob es sich bei den identifizierten Plateauphasen tatsächlich um Zeiträume mit konstantem Notenniveau handelt: In einem letzten Schritt werden die geglätteten Reihen und ihre zyklischen Komponenten z-standardisiert. Durch diese Standardisierung der Zeitreihen auf den gemeinsamen Erwartungswert 0 und die gemeinsame Varianz 1 lassen sich die Verlaufsformen unabhängig vom Niveau der Noten vergleichen. Die Grafiken (Abb.55-62) zeigen deutlich: Die sichtbaren Plateauphasen entstehen immer dort, wo relativ schwache Abwärtstrends mit den Aufwärtsbewegungen der Zyklen innerhalb der Zeitreihen aufeinandertreffen. Lediglich in Biologie und am deutlichsten in Psychologie überdauert die (letzte) Plateauphase die Aufwärtsbewegung der parallel verlaufenden Schwankung, was daran liegt, dass dort schon nahezu das untere Ende der Notenskala erreicht ist und der Abwärtstrend notwendigerweise auch ohne den lokalen Einfluss eines parallel verlaufenden Zyklus zum Erliegen kommt. Damit zeigen sich in den bisher untersuchten acht Studiengängen mit langfristiger Notenverbesserung zwei der theoretisch zu erwartenden Verlaufsformen gleichzeitig: Der Trend zu besseren Noten ebenso wie zyklische Schwankungen.

Abbildung 55: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. zyklische Komponente (LOWESS 0.3 – LOWESS 0.9, gestrichelt) BWL

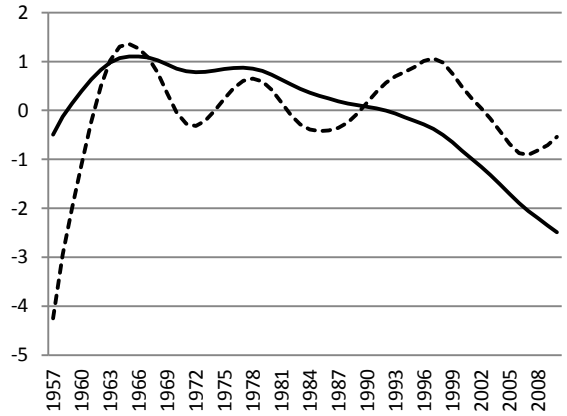


Abbildung 56: Durchschnittsnoten (LOWESS 0.2, durchgehend) vs. zyklische Komponente (LOWESS 0.2 – LOWESS 0.9, gestrichelt) VWL

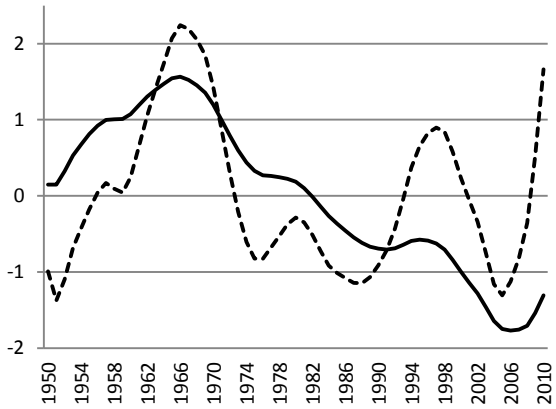


Abbildung 57: Durchschnittsnoten (LOWESS 0.4, durchgehend) vs. zyklische Komponente (LOWESS 0.4 – LOWESS 0.9, gestrichelt) Deutsch Lehramt

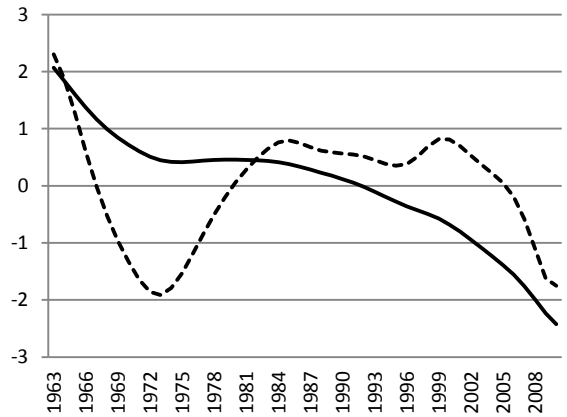


Abbildung 58: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. zyklische Komponente (LOWESS 0.3 – LOWESS 0.9, gestrichelt) Mathematik Lehramt

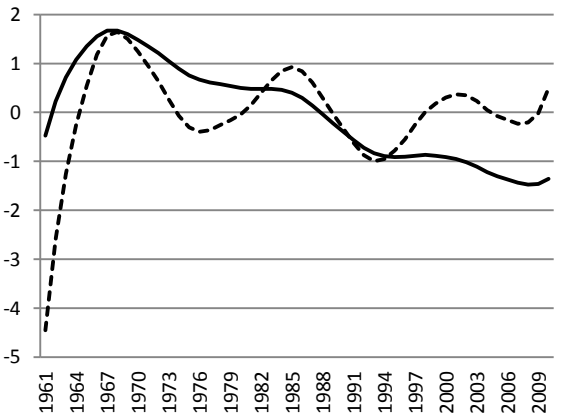


Abbildung 59: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. zyklische Komponente (LOWESS 0.3 – LOWESS 0.9, gestrichelt) Chemie

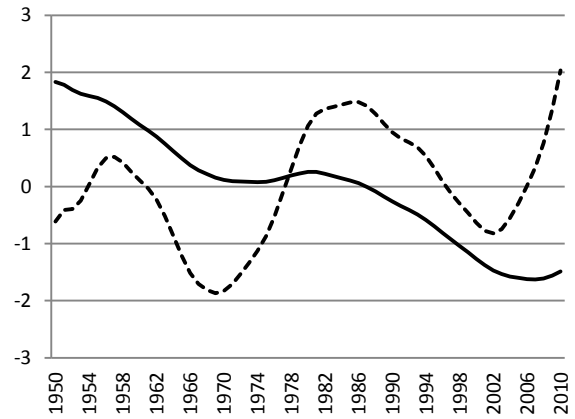


Abbildung 60: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. zyklische Komponente (LOWESS 0.3 – LOWESS 0.9, gestrichelt) Mathematik

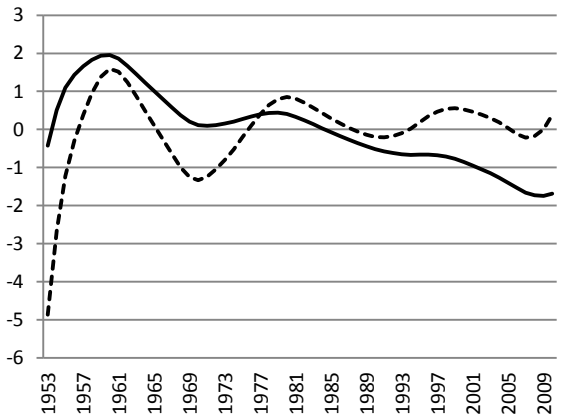


Abbildung 61: Durchschnittsnoten (LOWESS 0.4, durchgehend) vs. zyklische Komponente (LOWESS 0.4 – LOWESS 0.9, gestrichelt) Psychologie

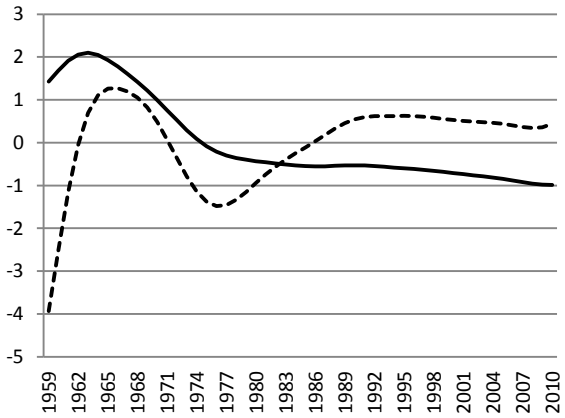
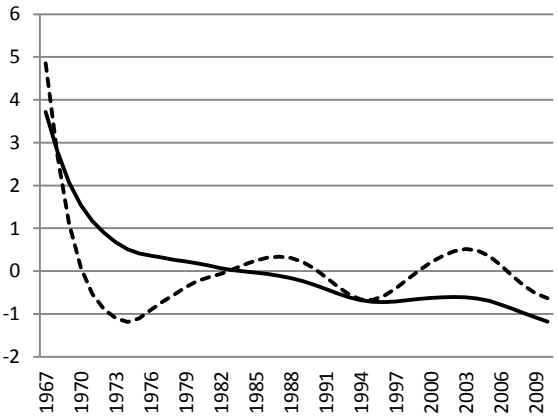


Abbildung 62: Durchschnittsnoten (LOWESS 0.4, durchgehend) vs. zyklische Komponente (LOWESS 0.4 – LOWESS 0.9, gestrichelt) Biologie



Dass die langfristige Verbesserung der Notendurchschnitte nicht erst ab einer bestimmten Annäherung an das Bestnotenniveau mit einer eingeschränkten Vergleichbarkeit der Leistungen einhergeht, belegen die Abbildungen 63-70[67]: In allen Studiengängen mit langfristiger Verbesserung sinkt parallel zum Notendurchschnitt auch die Streuung - unabhängig von der Höhe des Notenniveaus. In BWL sinkt die Standardabweichung im Zeitverlauf in etwa genauso stark wie in Chemie und Mathematik Diplom - trotz eines wesentlich schlechteren Notenniveaus, bei dem, im Gegensatz zu dem in Biologie und Psychologie, eine Verbesserung auch mit konstanter Streuung möglich wäre.

[67] Die Grafiken zeigen die mit dem Faktor 3 multiplizierten Werte der Standardabweichungen, die Koeffizienten und Gütemaße entstammen der Regression der nicht multiplizierten geglätteten Standardabweichungen auf die geglätteten Durchschnittsnoten. Die Koeffizienten sind in allen Fällen hochsignifikant (p=0.000).

Abbildung 63: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) BWL

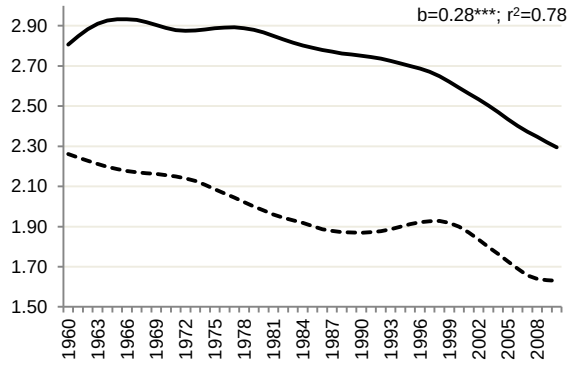


Abbildung 64: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) VWL

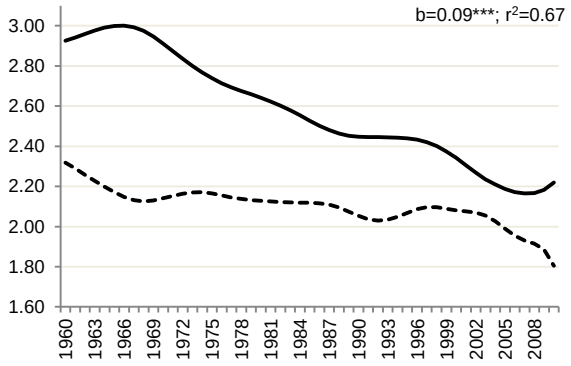


Abbildung 65: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Deutsch Lehramt

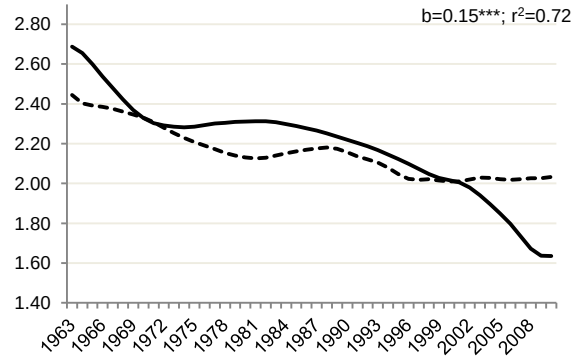


Abbildung 66: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Mathematik Lehramt

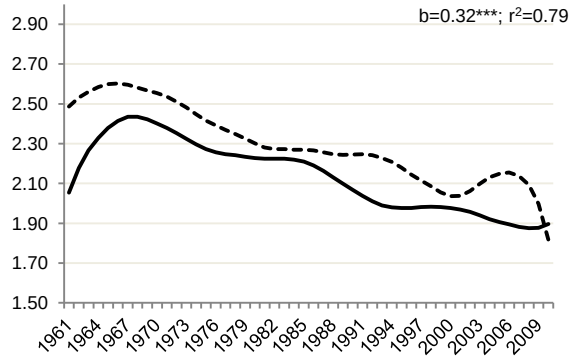


Abbildung 67: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Chemie

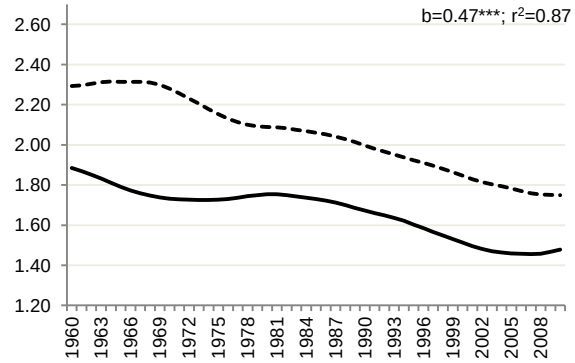


Abbildung 68: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Mathematik

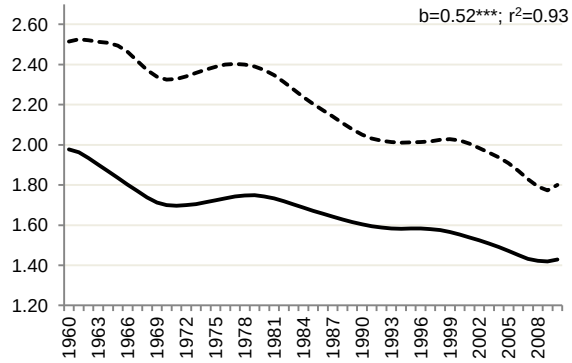


Abbildung 69: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Psychologie

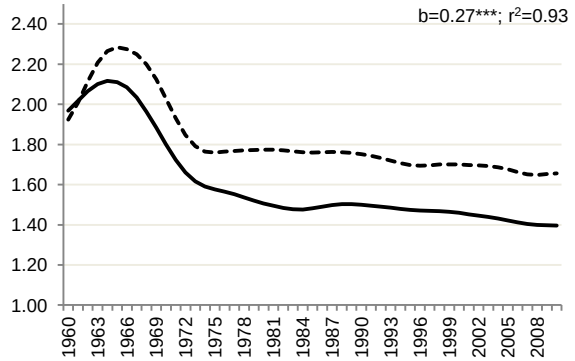
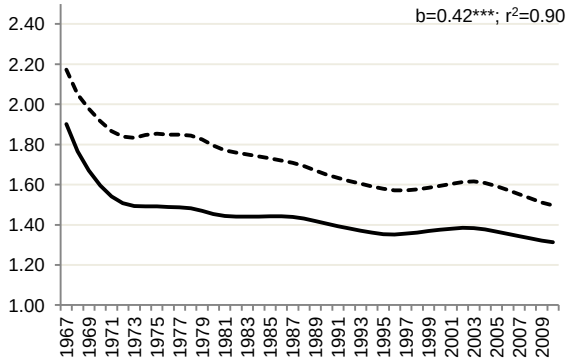


Abbildung 70: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Biologie



Wie die Aufarbeitung des Stands der Forschung im ersten Teil der Arbeit bereits vermuten ließ, sind nicht alle Studiengänge von einer langfristigen Verbesserung des Notenniveaus betroffen. Im Diplomstudiengang Maschinenbau, in den Magisterstudiengängen Soziologie und Germanistik sowie im ersten Staatsexamen der Rechtswissenschaften kann eine langfristig anhaltende Verbesserung nicht festgestellt werden (Abb.71 und 72). Während die Noten der ersten drei Studiengänge offensichtlich zyklisch verlaufen, scheinen die Noten in den juristischen Staatsexamen sich auf den ersten Blick konstant auf demselben Niveau zu bewegen. Der Eindruck eines linearen Verlaufs ist jedoch auf die grafische Skalierung zurückzuführen. Wird die Skalierung an die einzelne Reihe angepasst, offenbart sich auch für die scheinbar konstanten Noten der rechtswissenschaftlichen Prüfungen ein in wesentlich geringeren Spannweiten eingegrenzter, aber dennoch zyklischer Verlauf (Abb.73).

Die durchschnittlichen Abschlussnoten bewegen sich in Jura über den gesamten Zeitverlauf im Rahmen einer maximalen Spannweite von R=0.14 bei einer Standardabweichung von s=0.04, in Maschinenbau (R=0.28; s=0.08), Germanistik (R=0.54; s=0.12) und Soziologie (R=0.70; s=0.17) fallen diese Streuungswerte höher aus. Die Spannweite der Durchschnittsnoten in Jura liegt damit innerhalb des für die Studiengänge mit Verbesserung beobachteten Toleranzbereiches für die Plateauphasen. Zum letzten Messzeitpunkt 2007 liegt die Durchschnittsnote bei $\bar{x}$=3.30 und damit 0.08 Noten niedriger als zum ersten Messzeitpunkt 1959. Diese, auch aufgrund der hohen Fallzahlen hochsignifikante Veränderung (p=0.000) kommt dadurch zustande, dass der Wert 1959 den oberen, der Wert 2007 den unteren Wendepunkt der zyklischen Bewegung darstellt – ein weiterer Beleg dafür, wie leicht der Vergleich von zwei Zeitpunkten ohne Kenntnis der zwischenliegenden Entwicklung ein falsches Bild erzeugen kann.

Auch in Maschinenbau ist die Veränderung zwischen erstem und letztem Datenpunkt (-0.20) aus diesem Grund hochsignifikant (p=0.000). Im Vergleich zu den Verläufen in den anderen Studiengängen muss das *Notenniveau* in den Rechtswissenschaften dennoch - und auch trotz der zyklischen Verlaufsform - als über den Zeitverlauf konstant eingestuft werden. In Germanistik (-0.28; p=0.077) und Soziologie (+0.10; p=0.639) ist die Differenz zwischen Beginn und Ende der Zeitreihen nicht signifikant[68] und auch wenn die Noten in Germanistik ab Mitte der 1980er bis zum Ende der Zeitreihe sinken und somit für diesen Bereich ein Abwärtstrend festzustellen ist, wird hier die Position vertreten, dass diese Abwärtstendenz den sinkenden Part einer längeren zyklischen Bewegung darstellt, die sich daraus ergibt, dass der Zyklus in den vorliegenden empirischen Daten keinen idealtypischen symmetrischen Verlauf mit zentralen Peaks aufweist, sondern der Höhepunkt in den 1980ern Jahren um einige Jahre nach vorn verschoben auftritt.

[68] Im Studiengang Maschinenbau liegen nur Noten von zwei Hochschulen vor. In Soziologie und Germanistik sind zu Beginn der Zeitreihen zwei bzw. sechs Datenpunkte mit geringen Fallzahlen (n≤13 bzw. n≤10) entfernt worden.

Abbildung 71: Zeitlicher Verlauf der Abschlussnoten in Studiengängen ohne langfristige Notenverbesserung

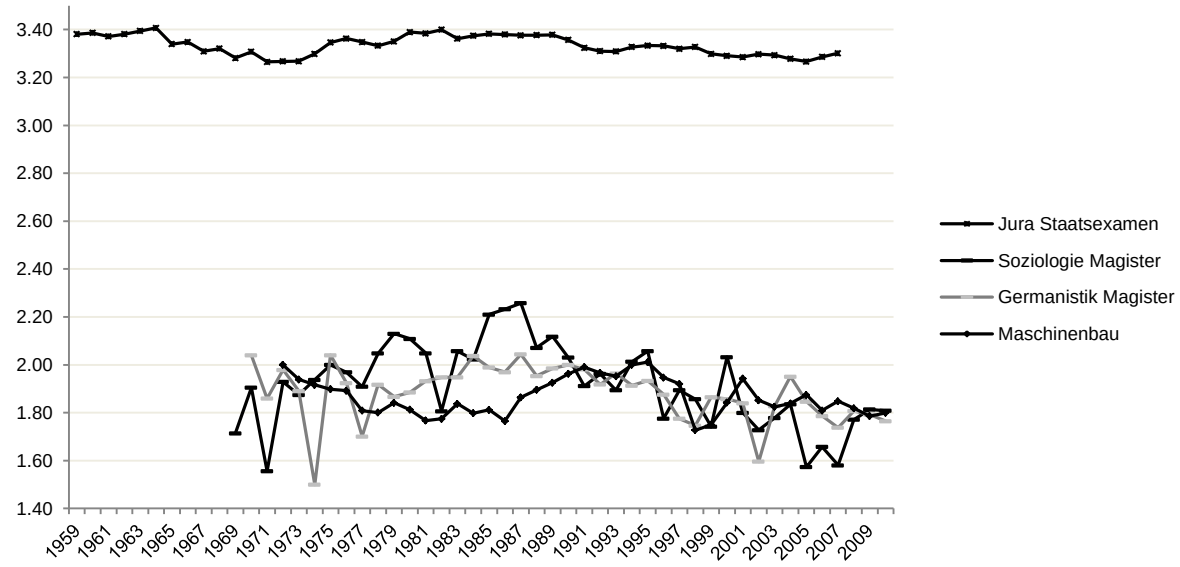


Abbildung 72: Zeitlicher Verlauf der Abschlussnoten in Studiengängen ohne langfristige Notenverbesserung (LOWESS 0.3)

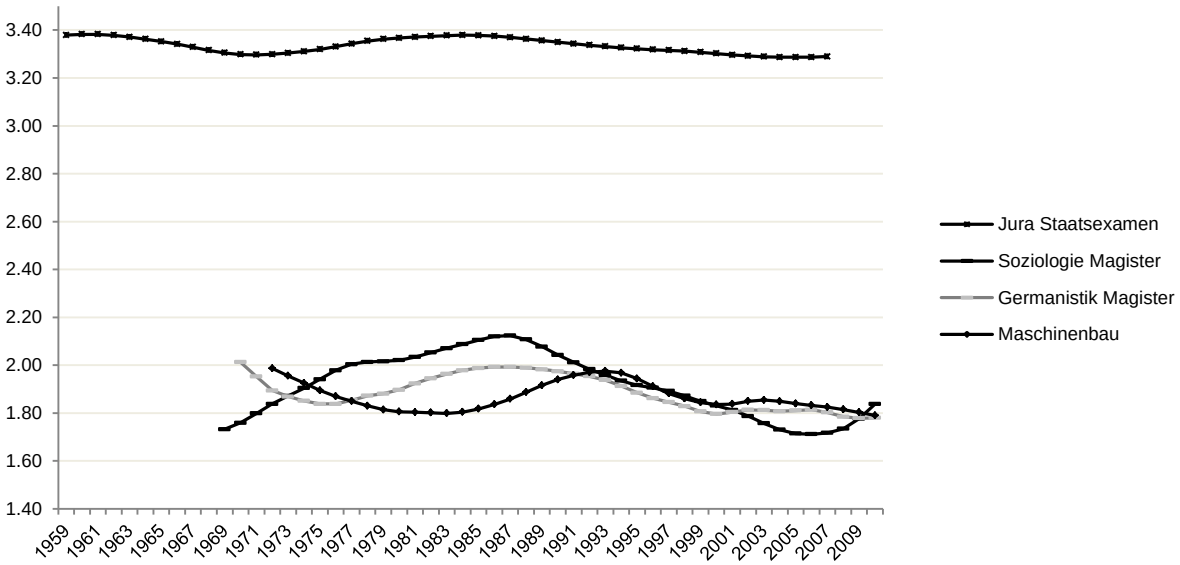
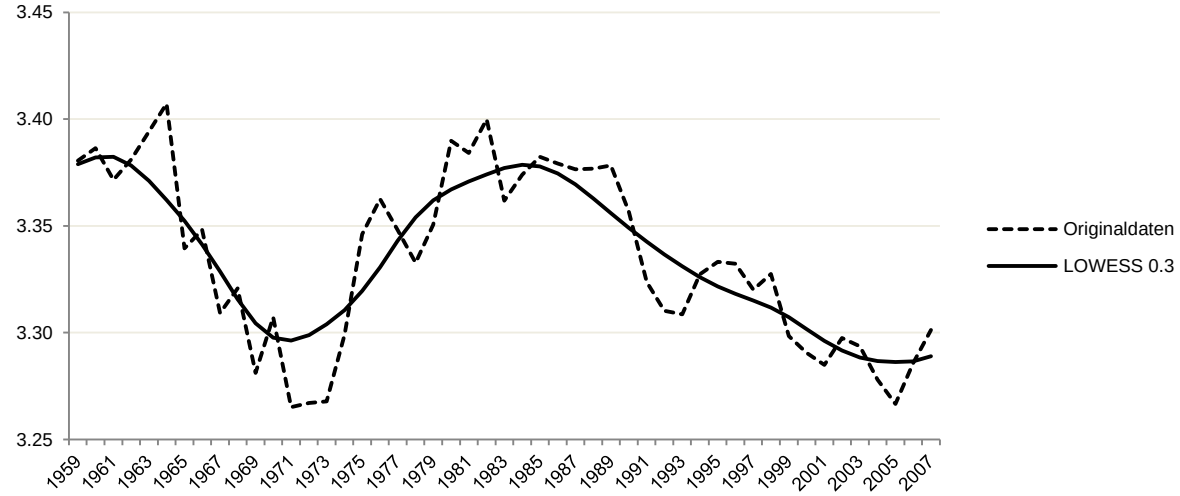


Abbildung 73: Zeitlicher Verlauf der Abschlussnoten im ersten juristischen Staatsexamen



Wie für die Gruppe der Studiengänge mit langfristiger Verbesserung zeigt sich auch an dieser Stelle, dass die einheitliche Verlaufsform der Noten über die Zeit nicht unbedingt mit einem einheitlichem Notenniveau und einem einheitlichem Start- und Endpunkt der einzelnen Verlaufsperioden einhergeht (Tab.17).

Die Noten in den beiden Magisterstudiengängen verlaufen mit nur geringen Unterschieden in der Notenhöhe relativ parallel. Bei genauer Betrachtung der reinen Verlaufsform zeigt sich jedoch, dass trotz gleichzeitigem Erreichen des Peaks die zyklische Bewegung der Noten in Soziologie wenige Jahre früher als in Germanistik beginnt und endet. In Maschinenbau setzt sie noch einmal später als in den Magisterstudiengängen ein und auch der Höhepunkt der zyklischen Bewegung tritt entsprechend verlagert auf. Die sich anschließende Abwärtsbewegung verläuft jedoch fast parallel, endet nur früher. Der Zyklus in Maschinenbau ist damit von kürzerer Dauer als in den anderen Studiengängen. In den juristischen Staatsexamina verläuft der Zyklus in etwa zeitglich mit der Notenbewegung in den Magisterstudiengängen, allerdings ist der Peak fünf Jahre früher erreicht, einhergehend mit einer steileren Aufwärts- und einer flacheren Abwärtsbewegung. Auffällig ist, dass die Zyklen in den Studiengängen ohne langfristige Verbesserung (außer in Maschinenbau) von ähnlicher Dauer sind wie die zyklischen Komponenten der Notenreihen der Studiengänge mit langfristiger Verbesserung.

Tabelle 17: Zyklusverlauf und Veränderungsausmaß in den Studiengängen ohne langfristige Notenverbesserung

Studiengang	Zyklusphase	Peak (schlechteste Note)	Maximale Spannweite (Wert/Jahre)	Spannweite in der Zyklusphase (Wert/Jahre)	Durchschnittliche Veränderung pro Jahr (Gesamt/ Zyklus)
Maschinenbau Diplom	1986-1998	1995 (2.01)	0.28/38	0.28/12	-0.005/-0.001
Germanistik Magister	1977-1998	1987 (2.04)	0.54/40	0.34/21	-0.007/+0.007
Soziologie Magister	1973-1996	1987 (2.26)	0.70/41	0.48/23	+0.002/+0.001
Jura Staatsexamen	1974-2000	1982 (3.40)	0.14/48	0.11/26	-0.002/-0.001

Wie die Abbildungen 74-77 zeigen, verlaufen die Streuungswerte auch in den Studiengängen ohne langfristige Verbesserung einigermaßen parallel zu den Durchschnittsnoten. Nur in Germanistik, wo das Notenniveau im erfassten Zeitraum aufgrund des Beginns der Reihe am oberen und des Endes am unteren Wendepunkt der zyklischen Bewegung sinkt, sinkt auch die Standardabweichung entsprechend. Dies belegt, dass die sinkende Streuung tatsächlich in Verbindung mit der Verbesserung im Zeitverlauf zu sehen ist, die in den acht Studiengängen mit Verbesserung zu beobachten ist und nicht eine generelle Tendenz der Notengebung darstellt. In den Rechtswissenschaften zeigt sich zudem äquivalent zur Entwicklung in Biologie und Psychologie, dass die Annäherung der Notendurchschnitte an die Grenzen der Notenskala zum Absinken der Streuung führt. Die höchsten Werte des Notenverlaufs treffen hier mit den niedrigsten Streuungswerten zusammen.

Abbildung 74: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*5 (LOWESS 0.3, gestrichelt) Jura

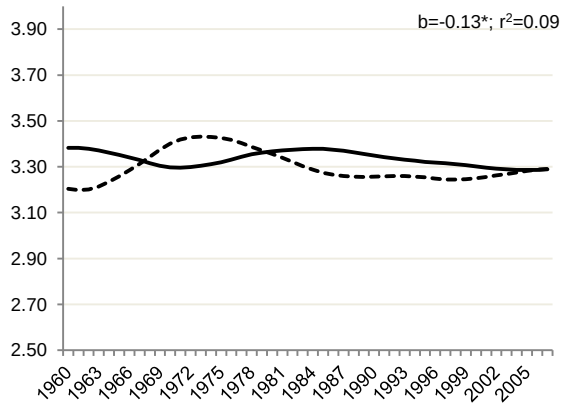


Abbildung 75: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Soziologie Magister

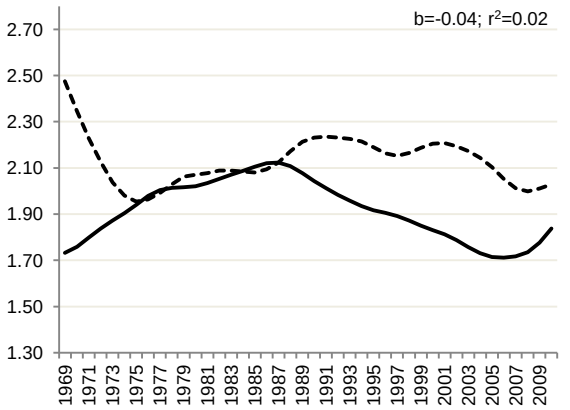


Abbildung 76: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Germanistik Magister

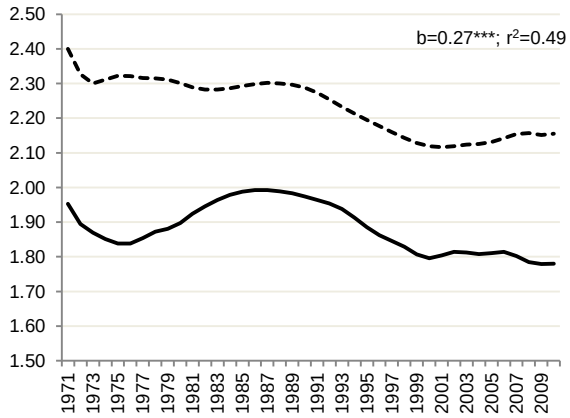
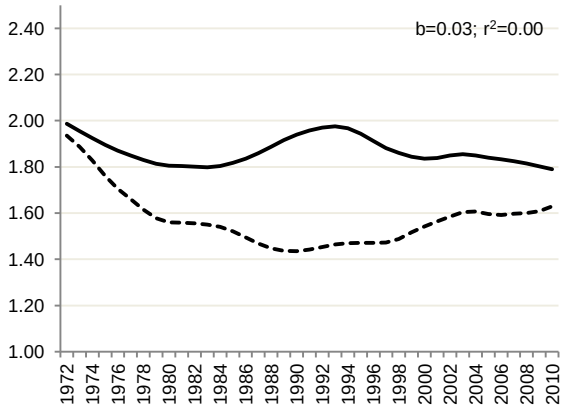
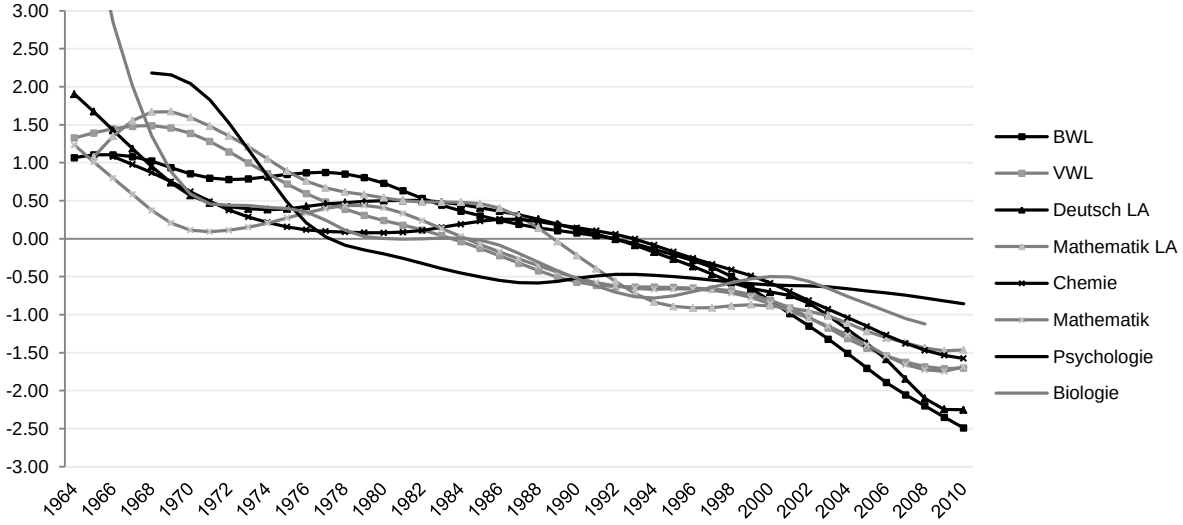


Abbildung 77: Durchschnittsnoten (LOWESS 0.3, durchgehend) vs. Standardabweichungen*3 (LOWESS 0.3, gestrichelt) Maschinenbau



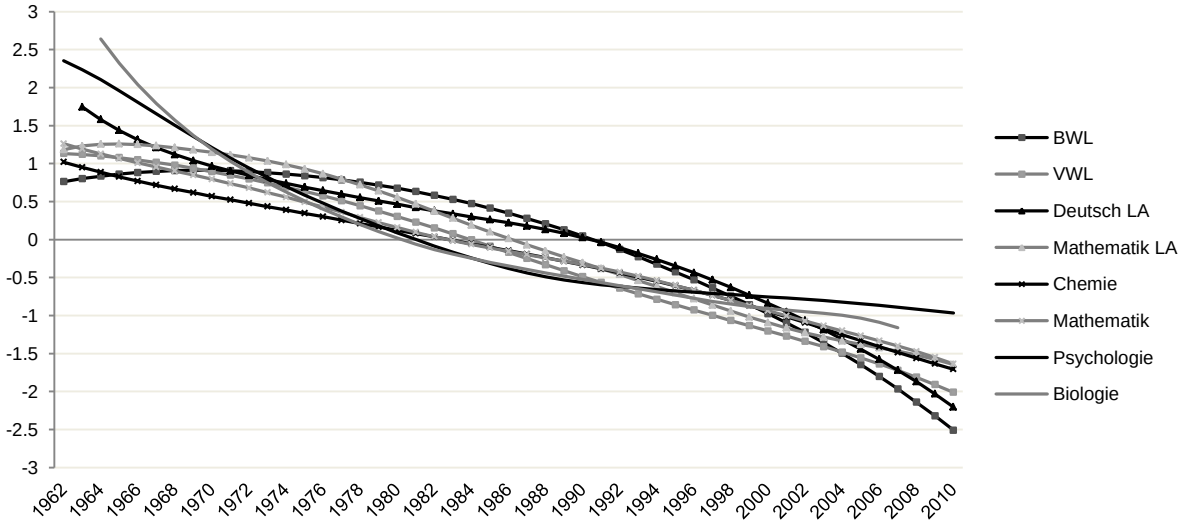
Besonders deutlich wird der Gegensatz Verbesserung vs. keine Verbesserung, wenn die geglätteten Daten z-standardisiert und zusätzlich im Zeitverlauf verschoben (englisch: gelagt) werden. Auf diese Weise wird sichtbar, dass sich die langfristige Entwicklung von Abschlussnoten auf Studiengangebene in nur zwei der zuvor beschriebenen vier idealtypischen Verlaufsformen (steigend, sinkend, konstant, zyklisch) einordnen lässt: Es sind einerseits (ab Anfang/Mitte der 1960er Jahre) sinkende (Abb.78 und 79) und andererseits zyklische verlaufende (Abb.80) Notenbewegungen zu beobachten, wobei die sinkenden Verläufe mehr oder weniger stark von zyklischen Bewegungen begleitet werden und nur annäherungsweise in ‚Reinform' auftreten. Dass die zyklischen Bewegungen der Studiengänge ohne langfristige Verbesserung den zyklischen Komponenten der Notenreihen mit Abwärtstrend ähneln, zeigt Abbildung 81. Hier ist schon eine deutlich stärkere zeitliche Verschiebung der Reihen nötig, um einen relativ übereinstimmenden Verlauf der standardisierten Reihen zu erhalten - dennoch ist das gleiche zyklische Muster erkennbar.

Abbildung 78: Z-Standardisierte, gelagte Zeitreihen der Abschlussnoten (LOWESS 0.3) in Studiengängen mit Verbesserung



Lags: VWL t+2, Mathematik LA t+1, Chemie t+6, Mathematik t+2, Psychologie t+4, Biologie t-2

Abbildung 79: Z-Standardisierte und gelagte Trendkomponenten in Studiengängen mit Verbesserung (LOWESS 0.9)

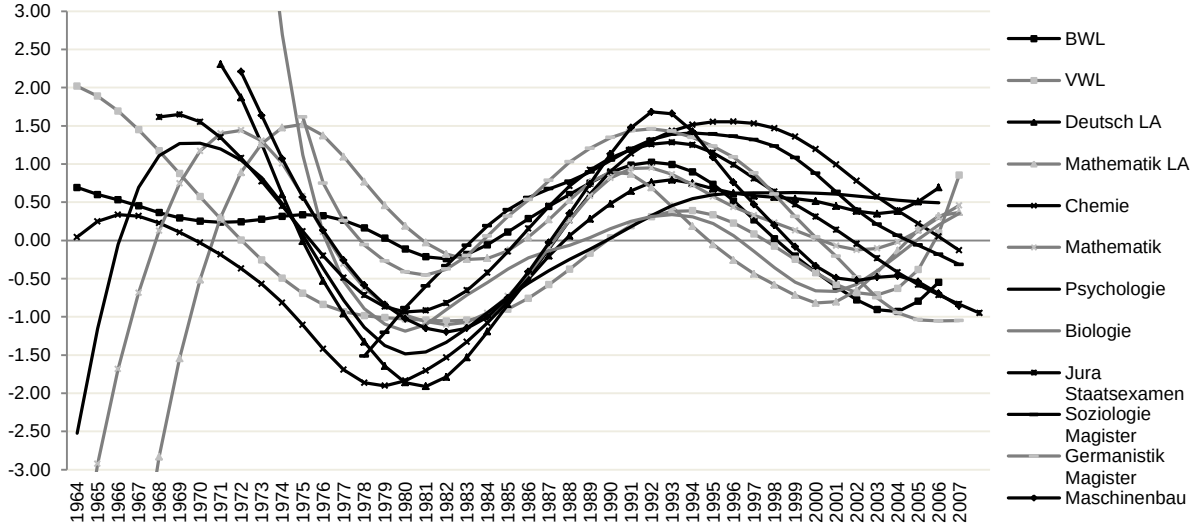


Lags: Chemie t+2, Mathematik t+2, Psychologie t+3, Biologie t-3

Abbildung 80: Z-Standardisierte, gelagte Reihen der Abschlussnoten (LOWESS 0.4) in Studiengängen ohne Verbesserung

Lags: Jura t+3, Soziologie t+1, Maschinenbau t-6

Abbildung 81: Z-Standardisierte, gelagte Reihen der Abschlussnoten (LOWESS 0.4) in Studiengängen ohne Verbesserung und der zyklischen Komponenten (LOWESS 0.4-LOWESS 0.9) der Abschlussnoten in den Studiengängen mit Verbesserung



Lags: BWL t-4, VWL t-3, Deutsch LA t+8, Mathematik LA t+6, Chemie t+10, Mathematik t+11, Psychologie t+4, Biologie t+6, Jura t+9, Soziologie Magister t+9, Germanistik Magister t+5

Es lässt sich festhalten, dass die durchschnittlichen Abschlussnoten sich langfristig in zwei unterschiedlichen Verlaufsformen entwickeln: Entweder sie verbessern sich langfristig, begleitet durch zyklische Schwankungen oder sie verlaufen zyklisch auf einem relativ stabilen Notenniveau. Die Diplomstudiengänge zählen mit Ausnahme von Maschinenbau alle zur ersten Gruppe, ebenso die beiden Lehramtsstudiengänge, die neben einem ähnlichen Notenniveau im Querschnitt auch im Längsschnitt eine zeitlich parallele Entwicklung aufweisen, wenn auch mit stärkerem Trend in Deutsch. Die beiden Magisterstudiengänge und das erste juristische Staatsexamen zählen zu Letzteren.

Auch für den Diplomstudiengang Maschinenbau kann keine langfristige Verbesserung festgestellt werden. Diesbezüglich muss jedoch auch bedacht werden, dass in Maschinenbau nur Daten von zwei Hochschulen vorliegen, weshalb von den vorliegenden Werten kein aussagekräftiger Rückschluss auf das im Studiengang üblicherweise vorzufindende Notenniveau und dessen Entwicklung gezogen werden kann[69].

Völlig konstante Notenniveaus sind über alle Prüflinge gemittelt nicht zu finden, auch langfristige Verschlechterungen der Noten sind nicht zu beobachten. In allen Studiengängen mit langfristiger Verbesserung setzt die erste Verbesserungsperiode in etwa zeitgleich zu Beginn/Mitte der 1960er Jahre ein und endet Anfang der 1970er Jahre, die zweite Verbesserungsphase beginnt jeweils im Laufe der 1980er Jahre, wobei sich das Ausmaß der Verbesserung und das Niveau auf dem sie sich vollzieht, studiengangspezifisch unterscheiden. Die Verbesserung der Noten wird durch Plateauphasen unterbrochen, die etwa zeitgleich einsetzen, allerdings von unterschiedlicher Dauer sind. Diese Plateauphasen treten immer dann auf, wenn der Abwärtstrend relativ schwach ist und im gleichen

[69] Grözinger (2017) kann für den Zeitraum von 1995 bis 2013 eine fast lineare Verbesserung für Maschinenbau an Universitäten nachweisen, wobei das Notenniveau, das er aus der Hochschulprüfungsstatistik berechnet über dem Mittelwert der beiden Hochschulen des hier verwendeten samples liegt.

Zeitraum eine Aufwärtsbewegung der zyklischen Zeitreihenkomponente, die in allen Studiengängen eine Dauer von ca. 20 Jahren aufweist, stattfindet.

Die isolierte Trendkomponente zeigt je nach Studiengang einen fast linearen, einen zu- oder abnehmenden Trend. In BWL und VWL, den Studiengängen mit dem schlechtesten Notenniveau der Studiengänge mit langfristiger Verbesserung nimmt die Stärke des Trends im Zeitverlauf zu. In Psychologie und Biologie ist die Verbesserung nach kurzer, starker Dynamik bereits nach der ersten Phase weitestgehend abgeschlossen, da die Noten kaum noch besser werden können, die Stärke des Trends nimmt entsprechend ab. Hier überdauert die Plateauphase die Aufwärtsphase der zyklischen Schwankung, seit Beginn der 1970er Jahre kann dort von grade compression gesprochen werden. Unabhängig vom studiengangspezifischen Notenniveau zeigt sich ein Zusammenhang zwischen dem Verlauf der Durchschnittsnoten und der zugehörigen Streuung: Mit dem Absinken des Notenniveaus verringert sich auch die Streuung, wodurch in den betroffenen Studiengängen die Differenzierung von Leistung zunehmend schwieriger geworden ist.

In den Studiengängen mit zyklischem Verlauf auf relativ stabilem Notenniveau verlaufen die Noten in unterschiedlichen Schwankungsbreiten, über die Jahre hinweg streuen sie am stärksten in Soziologie, am schwächsten in Jura. Die Spannweite dort beläuft sich mit R=0.13 auf einen Wert, der innerhalb des für die Studiengänge mit Verbesserung beobachteten Toleranzbereiches für die Plateauphasen liegt. Faktisch kann das Notenniveau dort trotz zyklischer Verlaufsform als über den Zeitverlauf konstant eingestuft werden. Die Auf- und Abwärtsbewegungen der Noten beginnen und enden leicht versetzt, die Zyklen dauern aber wie auch in den Studiengängen mit Verbesserung ca. 20 Jahre. Nur Maschinenbau fällt mit einer deutlich kürzeren Dauer aus dem Muster, was aber vermutlich auf die mangelnde Stichprobenbreite zurückzuführen ist. Sollte in Maschinenbau tatsächlich ebenfalls ein zyklischer Verlauf ohne Abwärtstrend vorliegen, gibt der Vergleich mit dem Notendurchschnitt für alle Hochschulen in Westdeutschland seit 1995 Grund zu der Annahme, dass die Dauer eines solchen Zyklus in den vorliegenden Daten deutlich unterschätzt wird. Ob überhaupt ein zyklischer Verlauf existiert, kann aber nicht mit Sicherheit gesagt werden.

Ansonsten bestätigt der Vergleich der Stichprobe mit den anderen verfügbaren Datenquellen die aus den einzelnen Datenpunkten grob ableitbare Tendenz zu besseren Noten seit den 1960er Jahren. Auch in der jährlichen Entwicklung seit 1995 zeigen sich keine nennenswerten Abweichungen in den Notenverläufen zwischen sample und (west-)deutschen Hochschulen. FH3a (an deutschen Hochschulen existieren hochschulübergreifend fach-/studiengangspezifische Notenverläufe) ist damit bestätigt.

8.1.3 Das Notenniveau und seine langfristige Entwicklung an den einzelnen Hochschulen

Da die Noten auf Studiengangebene als Mittel aller Prüflinge berechnet wurden, das Hochschulsample aber als Klumpenstichprobe konzipiert ist, kann das Notenniveau als Durchschnittwert nicht ein-

fach auf die einzelnen Hochschulen übertragen werden. Auch die Streuung der Noten aller Prüflinge hilft hier nur bedingt weiter, da einzelne Hochschulen auch systematisch geringer, andere stärker um den Durchschnitt streuen können. Da es zudem als gesichertes Wissen betrachtet werden kann, dass Hochschulunterschiede im Notenniveau auch innerhalb desselben Studiengangs ein verbreitetes Phänomen darstellen (Wissenschaftsrat 2003; 2007; 2012; Müller-Benedict/ Tsarouha 2011), das bisher jedoch in einer langfristigen Betrachtung noch nicht untersucht wurde, lohnt es sich, das Notenniveau an den einzelnen Hochschulen genauer zu betrachten.

Gleiches gilt für die langfristige Entwicklung. Hier ist der Verlauf der Noten auf Studiengangebene auch vom Verlauf an den Hochschulen mit den größeren Anteilen am gesamten Prüfungsvolumen abhängig. Die Noten müssen also auch in ihrer Entwicklungsform nicht zwangsläufig an allen Hochschulen gleichermaßen dem allgemeinen Trend entsprechen.

Die hochschulspezifische Analyse wird getrennt für die sechs Diplomstudiengänge Mathematik, Chemie, Biologie, VWL, BWL und Psychologie sowie für die Magister- und Lehramtsstudiengänge durchgeführt. Sie beschränkt sich auf diese Auswahl, da für Jura keine Daten auf Hochschulebene vorliegen, und die Analyse von nur zwei Hochschulen, für die Daten in Maschinenbau verfügbar sind, wenig aussagekräftig ist.

Anschließend wird überprüft, ob es studiengangübergreifende Muster der Notengebung an den einzelnen Hochschulen gibt, etwa ob an einer Hochschule in allen Studiengängen am besten bewertet wird, an einer anderen immer am schlechtesten. Zwar reichen die Daten für einzelne Hochschulen in einigen Studiengängen weiter zurück, jedoch werden die Analysen im Folgenden nicht vor dem Jahr 1960 beginnen. Dies geschieht zum einen aus dem einfachen Grund, dass die Datenaufnahme entsprechend des Projektfokus auf diesem Zeitraum in der Regel mit dem Jahr 1960 begann und die meisten Zeitreihen deshalb von 1960 bis 2010 reichen, zum anderen, um die Vergleichbarkeit innerhalb und zwischen den Studiengängen zu erhöhen.

Teilweise beginnen die Datenreihen einzelner Hochschulen oder auch aller Hochschulen in einzelnen Studiengängen erst einige Jahre nach 1960 oder weisen zu Beginn sehr niedrige Fallzahlen auf (v.a. Magister, Lehramt). Die Analyse beginnt dann entsprechend später (wenn alle Hochschulen betroffen sind) oder wird für zwei Zeiträume getrennt durchgeführt (einmal für den Zeitraum ab 1960, einmal für den Zeitraum, in dem für alle Hochschulen der Stichprobe Daten vorliegen).

Mathematik Diplom

Wie die grafische Darstellung[70] (Abb.82) bereits erkennen lässt, gibt es in Mathematik keine Hochschule im sample, deren Noten konstant völlig von den anderen abweichen. Immer wieder gibt es

[70] Aus Platzgründen werden an dieser Stelle nur die übersichtlicheren geglätteten Zeitreihen dargestellt. Grafiken der Originalzeitreihen finden sich im Anhang (Abb.A13-A22).

Überschneidungen zwischen mindestens zwei Hochschulen. Allerdings lassen sich mit Göttingen am oberen und Berlin sowie Heidelberg am unteren Ende Universitäten ausmachen, die über den größten Zeitraum die obere bzw. untere Begrenzung der Bandbreite darstellen, während etwa Karlsruhe konstant in der Mitte zwischen diesen Extremen liegt. Dies wird noch deutlicher in Abbildung 83, die die die Differenz zwischen durchschnittlicher Hochschulnote und Studiengangdurchschnitt, also dem Mittel aller Prüflinge ausweist. Werte >0 stehen für ein über dem Gesamtdurchschnitt liegendes Notenniveau, Werte<0 stellen ein im Vergleich zum Mittel aller Prüflinge unterdurchschnittliches Notenniveau dar. Die Differenzwerte zeigen, dass sich die Noten in Göttingen um die 0.2 Noten über, in Berlin und Heidelberg etwa 0.2 Noten unter dem Durchschnitt aller Prüflinge bewegen. Die Spannweite der maximalen Differenzen ist über den gesamten Zeitraum relativ konstant.

Abbildung 82: Durchschnittliche Abschlussnoten an den Hochschulen in Mathematik Diplom - Zeitverlauf (LOWESS 0.3)[71]

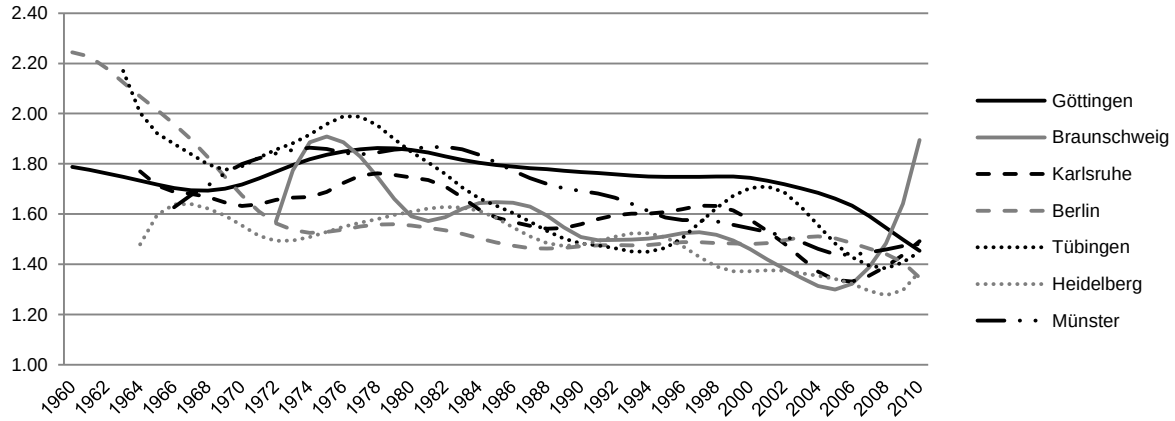
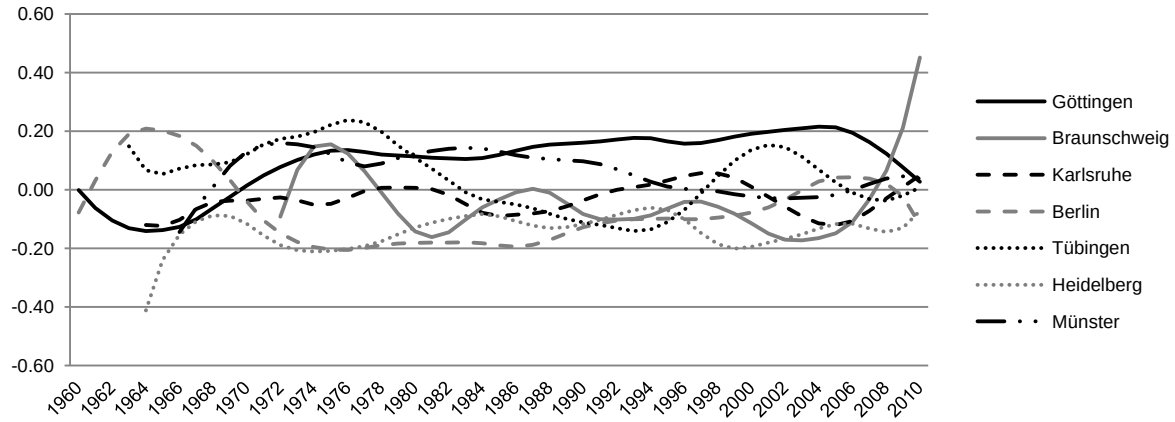


Abbildung 83: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Anhand der Originaldaten lässt sich die Differenz zwischen den Hochschulen genauer bestimmen: Im Mittel über den gesamten Zeitraum sind die Noten der Göttinger Prüflinge durchschnittlich 0.10 Noten schlechter als die aller Prüflinge, in Heidelberg sind sie durchschnittlich 0.13 Noten besser. Wird nur der Zeitraum betrachtet, in dem für alle Hochschulen im sample Werte vorliegen (1972-2010)

[71] In Braunschweig und Karlsruhe sind zu Beginn der Zeitreihen ein bzw. vier Datenpunkte mit geringen Fallzahlen (n=1 bzw. n<5) entfernt worden

liegen die Göttinger Noten 0.16 Noten über dem Durchschnitt, die Heidelberger 0.14 Noten darunter. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1960 gemittelt bei R=0.52 bzw. bei R=0.53 zwischen 1972-2010 und damit etwas höher als die geglätteten Daten es vermuten lassen. Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen über den gesamten Zeitraum gesehen 1961 (R=0.00) bzw. 1996 (R=0.29) für den Zeitraum in dem die Stichprobe komplett ist. Am größten ist diese Differenz 1965 (R=1.35) bzw. 2000 (R=0.91). Nach 1965 überschreitet sie den Wert R=0.91 nicht mehr und verläuft nahezu konstant.

Eine über den gesamten Zeitraum relativ konstante Notenhierarchie wie sie für die Studiengänge erstellt werden konnte, existiert nicht. Es lassen sich lediglich die in den meisten Jahren etwas zum Schlechteren bzw. zum Besseren abgesetzten Notenniveaus in Göttingen bzw. in Berlin und Heidelberg als Ansatz eines Musters festhalten.

Tabelle 18: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

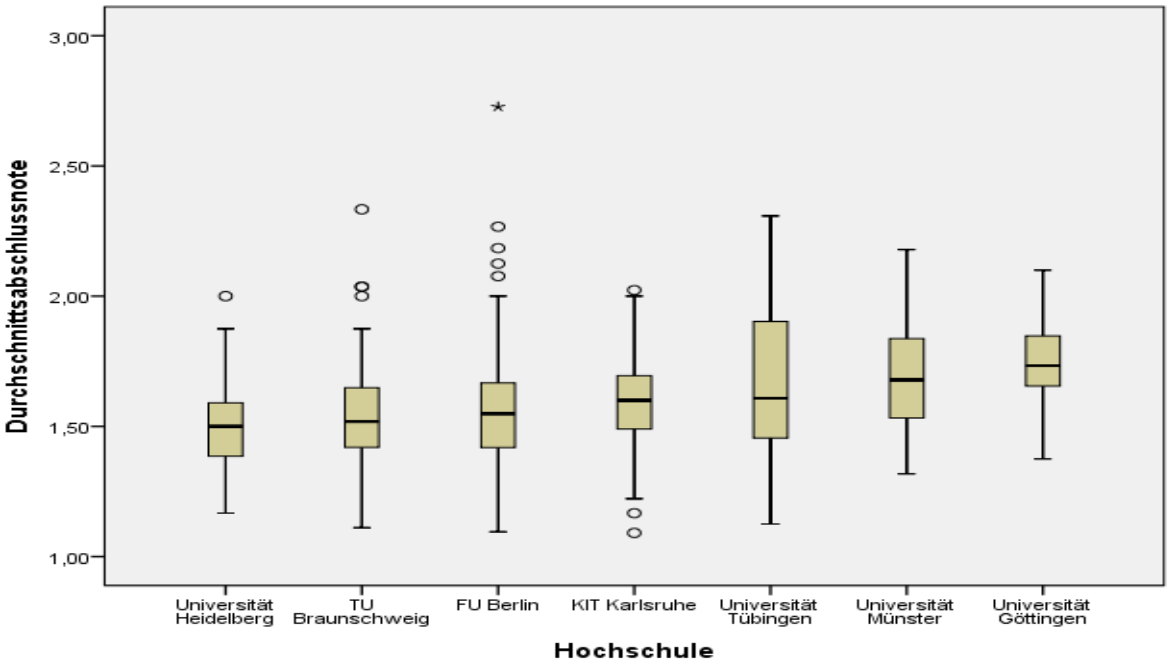
	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	HD	1.43	GÖ	1.64	HD	1.46	BER	1.55	BER	1.57	BER	1.36	TÜ	1.47	HD	1.38	BS	1.35	HD	1.25
2	KA	1.75	KA	1.66	BER	1.49	HD	1.58	BS	1.61	HD	1.42	BS	1.50	BER	1.44	KA	1.42	KA	1.37
3	GÖ	1.77	MS	1.71	KA	1.65	BS	1.74	KA	1.66	TÜ	1.48	BER	1.56	BS	1.53	HD	1.42	TÜ	1.38
4	TÜ	2.01	HD	1.72	GÖ	1.82	KA	1.79	HD	1.67	KA	1.51	HD	1.56	MS	1.58	MS	1.44	BS	1.46
5	BER	2.15	BER	1.73	BS	1.85	MS	1.84	TÜ	1.75	BS	1.58	KA	1.61	TÜ	1.60	BER	1.51	BER	1.47
6			TÜ	1.83	TÜ	1.85	GÖ	1.90	GÖ	1.80	MS	1.67	MS	1.66	KA	1.62	TÜ	1.63	MS	1.47
7					MS	1.85	TÜ	1.94	MS	1.93	GÖ	1.78	GÖ	1.76	GÖ	1.68	GÖ	1.81	GÖ	1.54

Eine einfaktorielle ANOVA zeigt, dass die Unterschiede zwischen den Hochschulen nur selten Signifikanz aufweisen: In gerade einmal 15 der 51 Jahre seit 1960 unterscheiden sich mindestens zwei Hochschulen in ihrer durchschnittlichen Abschlussnote signifikant voneinander. Da für einige Jahre Varianzhomogenität vorliegt, für andere nicht, wurden die Ergebnisse der ANOVA bis 1997 durch den Kruskal-Wallis Test überprüft und bestätigt. Lediglich ein Jahr, für das die ANOVA Signifikanz angibt, weist im Kruskal-Wallis Test keine Signifikanz auf, bei einem weiteren Jahr ist es genau umgekehrt. Da ab 1998 keine Individualdaten mehr vorliegen anhand derer die für den Kruskal-Wallis Test benötigten Rangreihen erstellt werden könnten, ist eine Anwendung dieses nicht-parametrischen Tests für die Jahre 1998-2010 nicht möglich.

Es lassen sich jedoch Post-Hoc Tests berechnen, so dass die Lücke anhand der Ergebnisse der Paarvergleiche geschlossen werden kann. Und deren Ergebnisse entsprechen ebenfalls weitestgehend den Ergebnissen der ANOVA. Der Games-Howell Test gibt für dieselben Jahre signifikante Differenzen aus, allerdings sind zwei Jahre mehr als signifikant ausgewiesen. Es unterscheiden sich die Noten in Göttingen und Berlin am häufigsten signifikant voneinander - aber auch nur in sechs von 51 Jahren. Mit durchschnittlich 27 bzw. 23 Prüflingen pro Jahr im Vergleichszeitraum sind die Fallzahlen zwar nicht besonders groß, es kann jedoch davon ausgegangen werden, dass die Differenzen im mittleren

Notenniveau nicht nur als Folge zu geringer Fallzahlen nicht signifikant sind - sozialwissenschaftliche Faustregeln variieren bezüglich einer aussagekräftigen Stichprobengröße für Paarvergleiche zwischen n>30 und n>50. Ein ähnliches Verhältnis besteht zwischen Heidelberg und Münster (fünf von 45 Jahren bei n=28 bzw. n=53). Die Abstände zwischen den einzelnen Hochschulen sind dabei sehr ähnlich: Über den gesamten Zeitraum gemittelt liegt die Distanz (also die Betragsfunktion der Differenz zwischen den Notenniveaus) im Bereich von 0.17 (zwischen Göttingen und Münster) bis 0.29 (zwischen Göttingen und Berlin sowie zwischen Göttingen und Heidelberg) Noten. Für 14 der 21 Paarvergleiche liegt dieser Wert im Bereich von 0.20 bis 0.25. Zwischen keinem Hochschulpaar besteht eine Periode durchgängig signifikanter Differenzen im hier verstandenen Umfang von mindestens drei aufeinanderfolgenden Jahren mit signifikanten Unterschieden im Notenniveau. Die größte prozentuale Differenz zwischen den Notenniveaus lässt sich über die Jahre gemittelt zwischen Heidelberg und Göttingen feststellen. Das Notenniveau in Heidelberg erreicht durchschnittlich 86.2% des Göttinger Niveaus (83.7% für den Zeitraum mit allen Hochschulen im sample). Entsprechend liegen diese beiden Hochschulen in der Verteilung der Abschlussnoten über die Zeit auch an den beiden Außenpolen der nach Höhe des Medians sortierten Darstellung der Boxplots:

Abbildung 84: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1960-2010



Es ist deutlich sichtbar, dass die durchschnittlichen Abschlussnoten in Tübingen über die Jahre mit Abstand am stärksten variieren, das heißt, die größte Spannweite einnehmen, abgesehen von den Ausreißern an der TU Braunschweig und der FU Berlin. In Göttingen und Heidelberg sind die in Relation hohen bzw. niedrigen Notenniveaus dagegen im Zeitverlauf am homogensten.

Werden die über die Jahre gemittelten Standardabweichungen und deren Standardabweichungen betrachtet, lässt sich erkennen, dass die Noten über die Zeit am stabilsten in Göttingen streuen. Die dortige Verteilung der Noten um den Mittelwert ist im Zeitverlauf deutlich homogener als an den

anderen Hochschulen, von denen fünf etwa eine doppelt so hohe Streuung der Streuung über die Zeit aufweisen (Spalten 4 und 5). Die Stärke der durchschnittlichen Streuung hingegen unterscheidet sich weniger stark, lediglich Münster fällt mit einem vergleichsweise leicht erhöhten Wert etwas aus der Reihe (Spalten 2 und 3). Auffällig ist, dass sich die mittlere Streuung in Berlin im Vergleichszeitraum ab 1972 (Spalte 3) deutlich gegenüber dem Zeitraum ab 1960 (Spalte 2) reduziert und damit den geringsten Wert für alle Hochschulen darstellt - und auch die Homogenität der Streuung im Zeitverlauf vergrößert sich für den kürzeren Vergleichszeitraum stark.

Tabelle 19: Streuung der Noten an den Hochschulen im Diplomstudiengang Mathematik

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1960-2010	1972-2010	1960-2010	1972-2010
Karlsruhe (ab 1964)	0.66	0.64	0.13	0.11
Heidelberg (ab 1964)	0.66	0.65	0.16	0.16
Braunschweig (ab 1972)	0.66	0.66	0.19	0.19
Berlin	0.68	0.62	0.18	0.13
Göttingen	0.72	0.72	0.09	0.09
Tübingen (ab 1963)	0.73	0.72	0.16	0.16
Münster (ab 1966)	0.77	0.75	0.17	0.15

Die Entwicklung der Noten an den einzelnen Hochschulen seit 1960 zeigt, dass in Göttingen ein relativ stabiles Niveau herrscht, das leichte Zyklen aufweist. Über den gesamten Zeitraum ist kaum eine Verbesserung des Notenniveaus festzustellen. Die maximale Verbesserung zwischen dem höchsten Wert in den 1960ern und dem niedrigsten im letztem Jahrzehnt beträgt zwar -0.61, wird zum Vergleich jedoch umgekehrt der niedrigste Wert in den 1960ern und der höchste Wert in den 2000er Jahren herangezogen, zeigt sich eine Verschlechterung von +0.68, was einmal mehr die eingeschränkte Aussagekraft des Vergleichs zwischen zwei Zeitpunkten unterstreicht.

Aussagekräftiger wird der Vergleich, wenn die Differenz zwischen dem jeweils über die ersten bzw. letzten 10 Jahre gemittelten Niveau herangezogen wird (Tab.20, Spalte 2): Von 1960 bis 1969 ergibt sich ein Wert von $\bar{x}$=1.72, der von 2001 bis 2010 mit $\bar{x}$=1.67 nur unwesentlich niedriger liegt (-0.05). In Berlin kann zunächst ein starker, ab Mitte der 1970er abgeschwächter und ab Mitte der 1980er Jahre nur noch relativ geringer Abwärtstrend beobachtet werden. Das Notenniveau liegt 2010 0.59 Noten unter dem Niveau von 1960 (Spalte 3) und ist in diesem Fall auch stellvertretend für die Veränderung des gemittelten Niveaus der 2000er Jahre gegenüber dem 1960er Jahrzehnt (Spalte 2). Die durchschnittliche jährliche Veränderung liegt bei -0.012 Noten pro Jahr über den gesamten Zeitraum (Spalte 4).

Wird statt dem gesamten Zeitraum nur der Zeitraum betrachtet, in dem der Trend am stärksten ist, im Folgenden immer verstanden als Zeitraum zwischen höchstem und niedrigstem Wert innerhalb der Abwärtsbewegung - in Berlin ist dies der Zeitraum zwischen 1965 und 2010 - beträgt die Differenz zwischen minimaler und maximaler Durchschnittsnote -1.63 Noten (Spalte 5). Die durchschnittliche jährliche Veränderung liegt in diesem Abschnitt der Reihe bei -0.036 (Spalte 6). Wird nur der Zeitraum ab 1972 betrachtet (Tab.21), zeigt sich, dass in Berlin in den 1970ern bereits annähernd das

mittlere Notenniveau der 2000er Jahre erreicht ist, während in Göttingen in den 1970ern das Höchstniveau herrscht, demgegenüber in den 2000ern eine Verbesserung erkennbar ist.

Die Noten in Münster weisen einen klaren Abwärtstrend auf, die Verbesserung verläuft hier am gleichmäßigsten, mit einer Trendstärke von -0.031 Noten zwischen 1981 und 2002, dem Zeitraum in dem sich die größte Verbesserung vollzieht. Auffällig ist, dass die Verbesserung erst gegen Anfang der 1980er Jahre einsetzt. In Braunschweig, Karlsruhe, Tübingen und Heidelberg verbessert sich das Notenniveau langfristig in viel deutlicheren Wellen. Die Niveauveränderung der 2000er gegenüber den 1960ern bzw. den 1970ern in Braunschweig ist an diesen Hochschulen vergleichbar, mit der in Münster (in Karlsruhe und Heidelberg etwas niedriger als diese), deutlich moderater als in Berlin und deutlich stärker als in Göttingen. Im stärksten Trendbereich ist die durchschnittliche Veränderung pro Jahr in Braunschweig, Tübingen und Münster jedoch ähnlich hoch wie in Berlin. Karlsruhe und Heidelberg weisen ebenfalls etwas niedrigere Werte und damit eine geringere maximale Abwärtsdynamik auf (Spalte 6), die sich zeitlich jedoch parallel zu der in Berlin vollzieht (Spalte 5), in Braunschweig und Tübingen erst 10 Jahre später einsetzt. Die Verbesserung überlagert in Braunschweig, Tübingen, Karlsruhe und Heidelberg erkennbar 10-20 jährige zyklische Bewegungen der Noten mit teils unterschiedlich starken Schwankungen, die eine Trendbereinigung noch besser sichtbar werden lässt (Abb.85). Bei genauerem Hinsehen lassen sich die Zyklen auch in den Berliner und Münsteraner Reihen finden, allerdings schwächer ausgeprägt.

Tabelle 20: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Mathematik 1960-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Berlin	-0.51	-0.59	-0.012	-1.63 (1965-2010)	-0.036
Braunschweig (ab 1972)	-0.38	+0.50	+0.013	-1.22 (1976-2007)	-0.039
Tübingen (ab 1963)	-0.33	-0.72	-0.015	-1.18 (1975-2007)	-0.036
Münster (ab 1966)	-0.33	-0.11	-0.002	-0.66 (1981-2002)	-0.031
Karlsruhe (ab 1964)	-0.25	-0.27	-0.006	-0.74 (1965-2004)	-0.019
Heidelberg (ab 1964)	-0.23	+0.07	+0.002	-0.82 (1966-2009)	-0.019
Göttingen	-0.05	-0.32	-0.006	--	--

Tabelle 21: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Mathematik 1972-2010

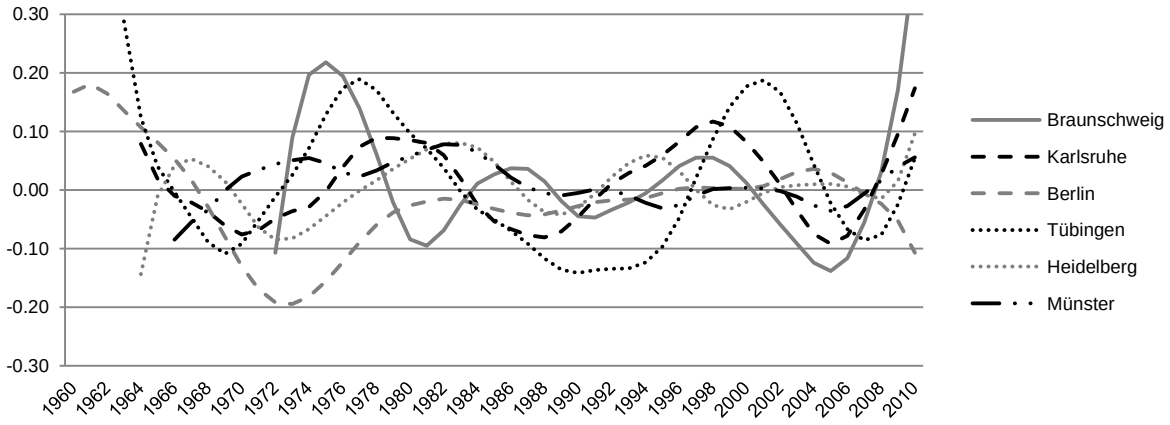
Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Münster	-0.43	-0.69	-0.018	-0.66 (1981-2002)	-0.031
Braunschweig	-0.38	+0.50	+0.013	-1.22 (1976-2007)	-0.039
Karlsruhe	-0.31	-0.10	-0.003	-0.62 (1979-2004)	-0.037
Heidelberg	-0.23	+0.07	+0.002	-0.68 (1983-2009)	-0.026
Göttingen	-0.21	-0.67	-0.017	--	--
Tübingen	-0.20	-0.40	-0.011	-1.18 (1975-2007)	-0.036
Berlin	-0.06	-0.29	-0.008	-0.76 (1991-2010)	-0.040

Die Betrachtung der Notenentwicklung an den einzelnen Hochschulen entspricht somit auch dem Verlauf der Noten auf Studiengangebene. An vier der sieben Universitäten (Braunschweig, Karlsruhe, Tübingen, Heidelberg) verläuft das Absinken des Notenniveaus in den Wellen, die sich auch über alle Prüflinge gemittelt zeigen. Zusätzlich ergibt sich durch die Kombination der Berliner und Münsteraner Zeitreihen mit kontinuierlicher Verbesserungstendenz bzw. mit nur sehr schwach ausgeprägten

Zyklen mit der Göttinger Zeitreihe, die ein konstantes Niveau mit Zyklen aufweist, ein ähnliches Muster. Ausgeprägte Plateauphasen lassen sich nicht erkennen, lediglich in Münster weisen die Noten über mehrere Jahre hinweg (1986-1995) eine relativ stabile Phase, das heißt, Schwankungen im Bereich <0.20 Noten, auf. Die Noten in Mathematik zeigen damit eine starke Dynamik an den Hochschulen und dies in der Mehrzahl sowohl im Trend als auch in der zyklischen Bewegung.

Die Noten in Berlin, Tübingen und Münster weisen in ihrem Verlauf die größten Veränderungen auf, wodurch sich auch die Differenzen zu den anderen Hochschulen und zum Durchschnitt aller Prüflinge verändern. Während die Noten in Berlin sich im Zeitverlauf immer schwächer verbessern und damit von unten dem Studiengangdurchschnitt annähern, ist für Münster ein gegenteiliger Effekt zu beobachten: Annäherung an das Mittel aller Noten durch relativ konstante und im Vergleich zu den anderen Hochschulen zunehmende Verbesserung ab Mitte der 1980er. In Tübingen führen die starken Zyklen zu Schwankungen um den gesamten Durchschnitt, so dass die Differenzen zu den einzelnen Hochschulen im Wechsel zu- und abnehmen.

Abbildung 85: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.3-LOWESS 0.9)



Mathematik Lehramt[72]

Im Lehramtsstudiengang Mathematik lassen sich auf den ersten Blick deutliche Differenzen zum Diplomstudiengang ausmachen (Abb.86 und 87): Die Spannweite zwischen bester und schlechtester Hochschule ist um ca. die Hälfte größer und es gibt mit Berlin eine Hochschule, die sich deutlich im Notenniveau von den übrigen unterscheidet. Ab Mitte der 1970er sind die Noten hier wesentlich schlechter als an den anderen vier Universitäten im sample, deren Noten sich über den größten Teil

[72] Anmerkungen zur Datenbasis: In Mathematik Lehramt stehen Daten für mindestens zwei Hochschulen ab 1961 und bis 2009 zur Verfügung. Drei Werte der Karlsruher Reihe von 2008-2010, die auf eine Lücke der Reihe von 1998-2007 (für diese Jahre liegen in der Prüfungsstatistik keine Informationen vor) folgen, werden nicht berücksichtigt. Für Tübingen sind ebenfalls nur Daten bis 1997 vorhanden, für Braunschweig bis 2008 und für Berlin bis 2009. Die Durchschnittsnoten in Karlsruhe, Berlin und Tübingen enthalten die Noten aller im jeweiligen Landesprüfungsamt bestandenen Prüfungen, welche anhand der Zeugnisse nicht mehr eindeutig einzelnen Hochschulen zuzuordnen waren. So sind in den Karlsruher Noten auch die Absolvent*innen aus Mannheim und Heidelberg enthalten, in den Tübinger Noten die aus Ulm und in den Berliner Noten die, aller an Berliner Hochschulen abgelegten Lehramtsprüfungen.

der beobachteten Jahre in ihrer Höhe ähneln. Die Spannweite nimmt mit der Zeit geringfügig ab, was aber vor allem auf die große Differenz im Niveau zwischen Göttingen und Karlsruhe bzw. Braunschweig zu Beginn der Reihe zurückzuführen ist.

Abbildung 86: Durchschnittliche Abschlussnoten an den Hochschulen in Mathematik Lehramt - Zeitverlauf (LOWESS 0.3)

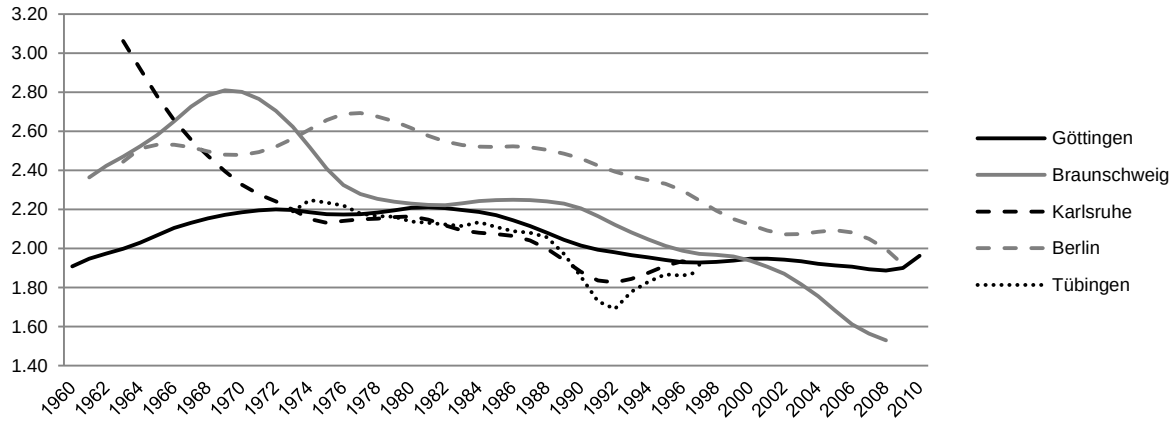
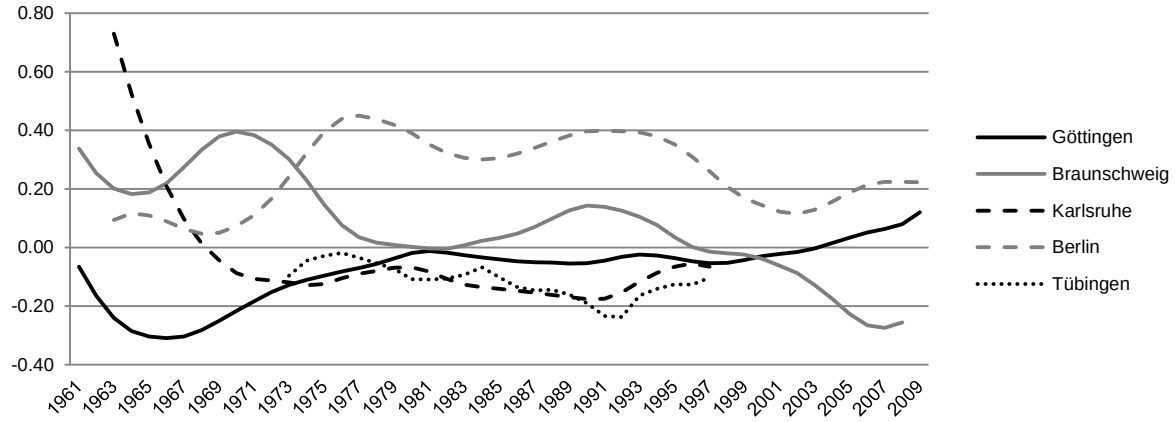


Abbildung 87: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Die Originaldaten zeigen, dass die Noten in Tübingen im Durchschnitt 0.11 Noten unter dem Mittel aller Prüflinge liegen und damit die besten Ergebnisse darstellen, während die Berliner Noten über den gesamten Zeitraum betrachtet 0.26 Noten schlechter als der Studiengangdurchschnitt sind. Im Zeitraum von 1973 bis 1997, in dem für alle fünf Hochschulen vollständige Daten vorliegen, sind die Berliner Noten im Durchschnitt sogar 0.38 Noten schlechter als die aller Lehramtsabsolvent*innen in Mathematik, die besten Noten in diesem Zeitraum gibt es in Karlsruhe (-0.12 im Vergleich zum Studiengangniveau). Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1961 gemittelt bei R=0.60 bzw. bei R=0.61 zwischen 1973-1997 und damit nicht so stark über den Werten im Diplomstudiengang, wie die grafische Darstellung nahelegt.

Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen über den gesamten Zeitraum gesehen 2000 (R=0.13) bzw. 1995 (R=0.29) für den Zeitraum in dem die Stichprobe komplett ist. Am größten ist diese Differenz 2007 (R=1.16) bzw. 1990

(R=0.81). Trotz des Maximums 2007 nimmt die maximale Differenz in den Noten zwischen den Hochschulen im Zeitverlauf tendenziell ab.

Über jeweils 5-Jahresabschnitte gemittelt zeigt sich von ca. 1976-2000 eine grobe Notenhierarchie: Die besten Noten gibt es entweder in Karlsruhe oder Tübingen. Zunächst über Göttingen, dann über Braunschweig werden sie schlechter, in Berlin sind sie am schlechtesten.

Tabelle 22: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

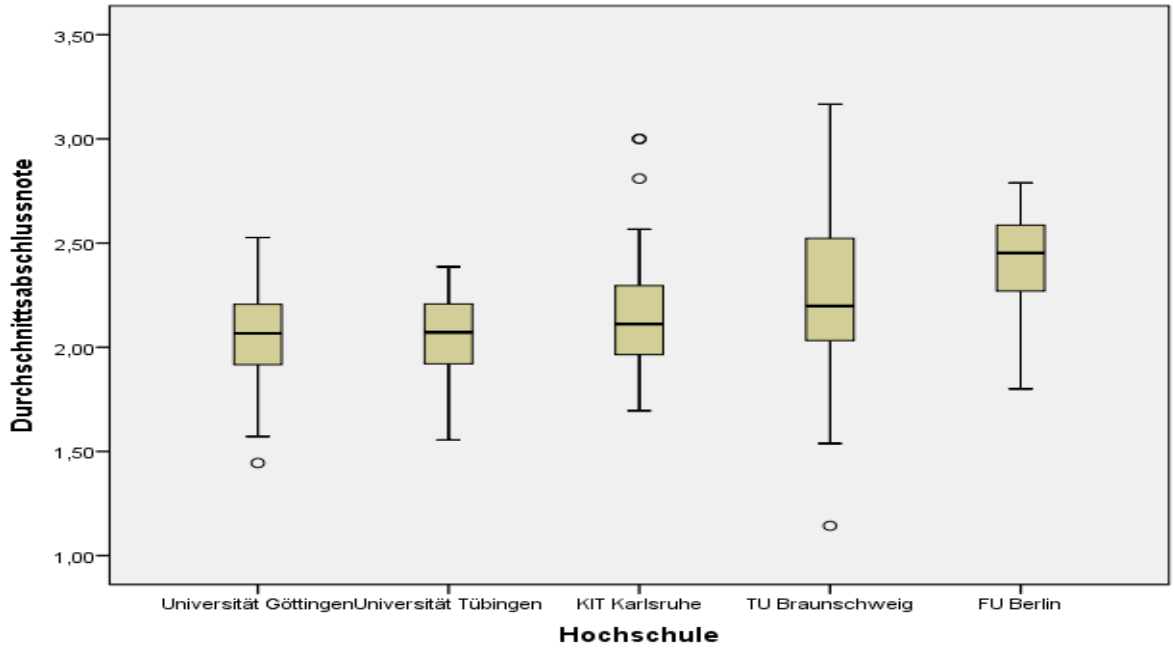
	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	GÖ	2.05	GÖ	2.13	GÖ	2.17	KA	2.12	KA	2.08	KA	2.02	TÜ	1.77	TÜ	1.83	BS	1.86	BS	1.55
2	BS	2.47	KA	2.45	KA	2.20	TÜ	2.16	TÜ	2.12	TÜ	2.02	KA	1.85	KA	1.91	GÖ	2.02	GÖ	1.68
3	BER	2.52	BER	2.51	TÜ	2.27	GÖ	2.20	BS	2.21	GÖ	2.09	GÖ	1.91	GÖ	1.94	BER	2.07	BER	2.05
4	KA	2.94	BS	2.80	BER	2.53	BS	2.21	GÖ	2.23	BS	2.30	BS	2.05	BS	1.96				
5					BS	2.64	BER	2.72	BER	2.47	BER	2.57	BER	2.32	BER	2.23				

Die einfaktorielle ANOVA zeigt, dass die Unterschiede zwischen den Hochschulen häufiger Signifikanz aufweisen als im Diplom: In 26 der 49 Jahre seit 1961 unterscheiden sich mindestens zwei Hochschulen in ihrer durchschnittlichen Abschlussnote signifikant voneinander. Da auch in den Lehramtsdaten für einige Jahre Varianzhomogenität vorliegt, für andere nicht, wurden die Ergebnisse der ANOVA bis 1997 wieder durch den Kruskal-Wallis Test überprüft und bestätigt. Lediglich ein Jahr, für das die ANOVA Signifikanz angibt, weist im Kruskal-Wallis Test keine Signifikanz auf. Der Games-Howell Test, wieder berechnet für die Jahre ab 1998, gibt ebenfalls für ein Jahr weniger signifikante Differenzen aus, so dass der nicht-parametrische Test 24 statt 26 signifikante Unterschiede ausweist.

Es unterscheiden sich die Noten in Tübingen und Berlin am häufigsten signifikant voneinander - in 10 von 25 Jahren. Mit durchschnittlich 49 bzw. 32 Prüflingen pro Jahr im Vergleichszeitraum sind die Fallzahlen etwas größer als im Diplom und damit ausreichend für die durchgeführten Tests. Perioden durchgängig signifikanter Unterschiede im Notenniveau existieren nur im Vergleich der Berliner Noten mit denen in Göttingen, Tübingen und Karlsruhe, was auf die zum Schlechteren abgesetzte Position der Berliner Noten zu den meisten Zeitpunkten zurückzuführen ist, sind aber auch nur von vier bis sechs Jahren Dauer. Die Distanzen (Betragsfunktion der Differenz) zwischen den einzelnen Hochschulen reichen von 0.14 zwischen Karlsruhe und Tübingen bis 0.50 zwischen Berlin und Tübingen. Fünf der 10 Paarvergleiche liegen im Bereich von 0.22 bis 0.33 Noten Distanz.

Tendenziell sind die Abstände im Notenniveau zwischen den Hochschulen damit etwas höher als im Diplomstudiengang Mathematik. Die größte prozentuale Differenz zwischen den Notenniveaus lässt sich über die Jahre gemittelt zwischen Tübingen und Berlin feststellen: Das Notenniveau in Tübingen erreicht durchschnittlich 80.35% des Berliner Niveaus (für den Zeitraum mit allen Hochschulen im sample liegt der Anteil des Karlsruher Niveaus mit 80.34% am Berliner Niveau noch minimal niedriger). Über alle Jahre betrachtet, streut das Notenniveau in Braunschweig wesentlich breiter als es an den anderen Standorten der Fall ist, die, von den Ausreißern abgesehen, eine ähnliche breite Spannweite an Durchschnittsnoten abdecken.

Abbildung 88: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1961-2009



Die über die Jahre gemittelten Standardabweichungen und deren Standardabweichungen zeigen, dass die Noten über die Zeit am stabilsten in Tübingen streuen - allerdings sind dort auch weniger Jahre berücksichtigt. Die dortige Verteilung der Noten um den Mittelwert ist im Zeitverlauf deutlich homogener als an den anderen Hochschulen, von denen drei eine mehr als doppelt so hohe Streuung der Streuung über die Zeit aufweisen (Spalten 4 und 5). Die Stärke der durchschnittlichen Streuung hingegen unterscheidet sich weniger stark, lediglich Göttingen fällt mit einem vergleichsweise niedrigen Wert etwas aus der Reihe (Spalten 2 und 3). Zwischen den beiden untersuchten Vergleichszeiträumen gibt es keine großen Unterschiede, die mittlere Streuung fällt im kürzeren Zeitraum geringfügig niedriger aus. Im Vergleich zum Diplomstudiengang liegt die mittlere Streuung über alle Hochschulen betrachtet etwas höher.

Tabelle 23: Streuung der Noten an den Hochschulen im Lehramtsstudiengang Mathematik

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1961-2009	1973-1997	1961-2009	1973-1997
Göttingen	0.64	0.63	0.16	0.18
Berlin (1963-2009)	0.73	0.70	0.15	0.13
Braunschweig (bis 2008)	0.75	0.71	0.16	0.13
Tübingen (1973-1997)	0.77	0.77	0.07	0.07
Karlsruhe (1963-1997)	0.81	0.78	0.10	0.07

Wird die Entwicklung der Noten an den einzelnen Hochschulen seit 1961 im Verlauf betrachtet, so zeigt sich in Göttingen eine lange Wellenbewegung, die gegen Ende der Reihe eine Drittelnote unter dem Niveau von 1961 liegt. In Braunschweig, Karlsruhe und Berlin verbessert sich das Notenniveau (in dieser Reihenfolge abnehmend) deutlich stärker und langfristig in Wellen. Hier überlagert die unterschiedlich starke Verbesserung wiederum erkennbar 10-20 jährige zyklische Bewegungen der Noten mit teils unterschiedlich starken Schwankungen - eine Trendbereinigung macht dies auch hier deutlicher sichtbar (Abb.89). Im stärksten Trendbereich ist eine Bandbreite von -0.035 Noten Verbes-

serung pro Jahr in Göttingen bis zu -0.056 Noten pro Jahr in Braunschweig vorhanden. Zeitlich umschließt dieser Bereich in allen Fällen die 1980er Jahre, der Anfangspunkt des stärksten Trendbereichs variiert jedoch von 1964 in Karlsruhe bis 1983 in Göttingen und auch das Ende liegt in einem weiten Zeitraum zwischen 1991 in Tübingen und 2009 in Göttingen verteilt.

In Braunschweig und Tübingen vollzieht sich die Abwärtsdynamik in ihrer kraftvollsten Phase am stärksten, Karlsruhe und Berlin weisen noch eine stärkere maximale Abwärtsbewegung auf als Göttingen. In Tübingen ist eine ähnlich starke Verbesserung von erstem zum letzten Messzeitpunkt wie in Göttingen im beobachteten Zeitraum von 1973-1997 zu verzeichnen, allerdings ist hier nur die Andeutung eines Zyklus erkennbar, was möglicherweise auf die Kürze der Zeitreihe zurückzuführen ist. Nur für diesen Zeitraum betrachtet, gleichen sich die absoluten Veränderungen mit Ausnahme von Braunschweig gegenüber dem Gesamtzeitraum erheblich an, die durchschnittlichen Veränderungen der Noten pro Jahr im stärksten Trendbereich bleiben allerdings in einem ähnlichen Wertebereich.

Tabelle 24: Kennzahlen - Notenentwicklung an den Hochschulen im Lehramtsstudiengang Mathematik 1961-2009

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Braunschweig (bis 2008)	-0.86	-0.75	-0.016	-2.03 (1971-2007)	-0.056
Karlsruhe (1963-1997)	-0.67	-1.12	-0.033	-1.30 (1964-1992)	-0.047
Berlin (ab 1963)	-0.43	-0.42	-0.009	-0.99 (1976-2000)	-0.041
Tübingen (1973-1997)	-0.33	-0.29	-0.012	-0.83 (1975-1991)	-0.052
Göttingen	-0.23	-0.36	-0.007	-0.93 (1983-2009)	-0.035

Tabelle 25: Kennzahlen - Notenentwicklung an den Hochschulen im Lehramtsstudiengang Mathematik 1973-1997

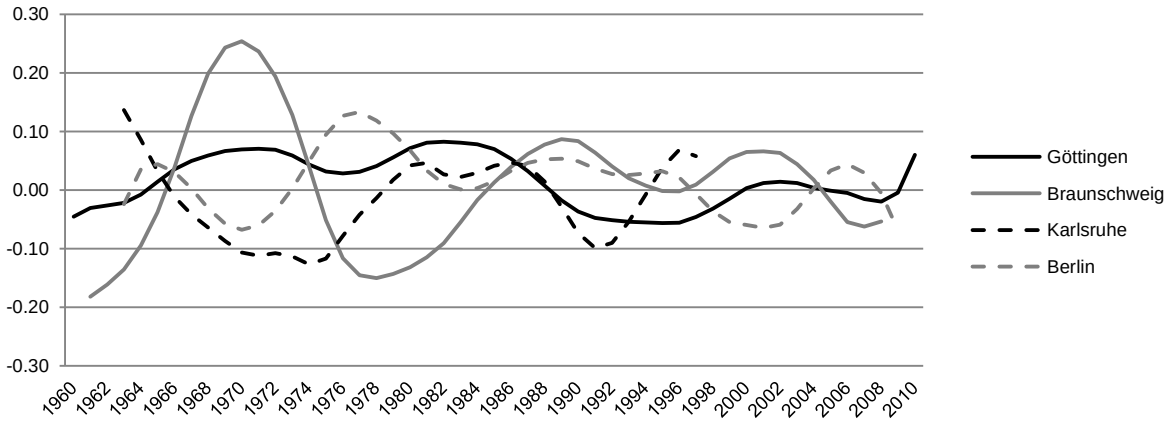
Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Tübingen	-0.33	-0.29	-0.012	-0.83 (1975-1991)	-0.052
Karlsruhe	-0.29	-0.18	-0.007	-0.60 (1979-1992)	-0.046
Berlin	-0.26	-0.28	-0.012	-0.66 (1976-1995)	-0.035
Braunschweig	-0.20	-0.88	-0.037	-0.88 (1973-1997)	-0.037
Göttingen	-0.18	-0.26	-0.011	-0.55 (1983-1994)	-0.050

Die Betrachtung der Notenentwicklung an den einzelnen Hochschulen entspricht auch dem Verlauf der Noten auf Studiengangebene. An vier der fünf Universitäten (in Tübingen nur ansatzweise) verlaufen die Noten in Wellen – ebenfalls in vier von fünf Hochschulen (Ausnahme: Göttingen in langer Wellenform) verbessern sich die Noten nahezu durchgängig. Plateauphasen lassen sich Mitte der 1970er/Anfang der 1980er in Berlin (1973-1980), Braunschweig (1979-1984) und Tübingen (1976-1982) erkennen. Die Noten in Mathematik Lehramt zeigen insgesamt ebenfalls eine starke Dynamik an den Hochschulen, sowohl im Trend als auch in der zyklischen Bewegung.

Die Noten in Braunschweig und Göttingen weisen im Verhältnis zum Studiengangmittel die größten Veränderungen im Zeitverlauf auf: Die Göttinger Noten liegen zunächst deutlich unter dem Durchschnitt, näheren sich dann zunehmend an und liegen am Ende des Beobachtungszeitraums leicht über dem Durchschnitt aller Prüflinge. Die Braunschweiger Noten nehmen den umgekehrten Verlauf.

In Tübingen gibt es für den kurzen Zeitraum relativ konstant Noten unter dem Studiengangdurchschnitt, in Berlin zu Beginn und Ende im Vergleich weniger schlechte Noten als in der Mitte der Reihe.

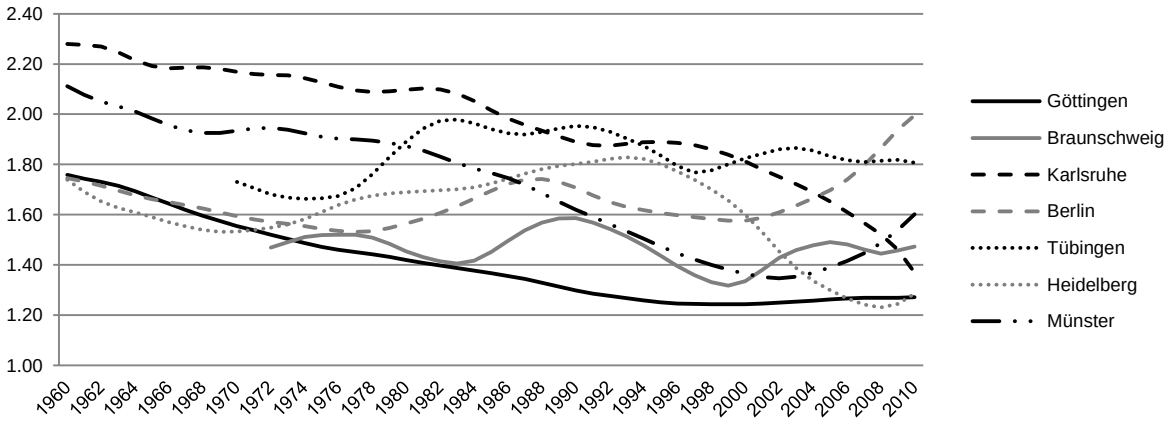
Abbildung 89: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.3-LOWESS 0.9)



Chemie Diplom

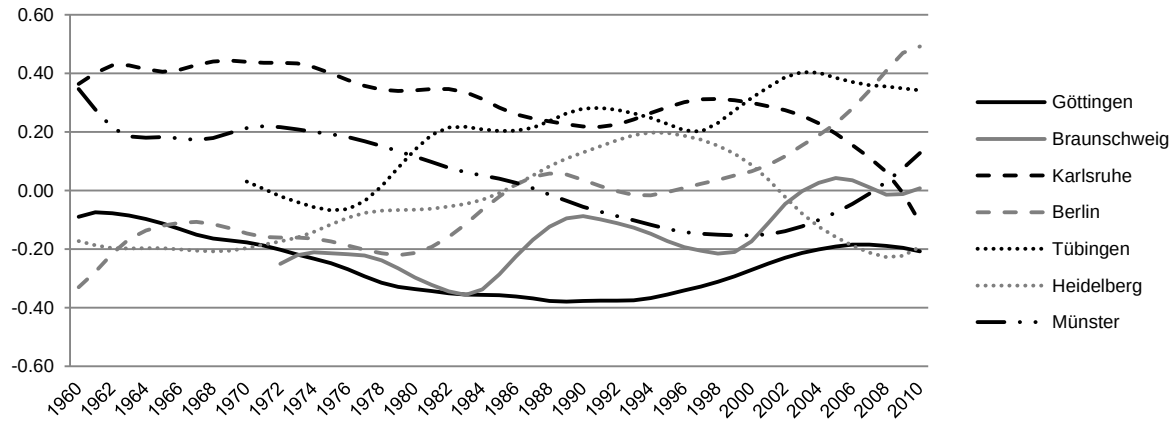
Es fällt anhand der grafischen Analyse auf, dass die Bandbreite der Durchschnittsnoten in Chemie etwas größer ist als in Mathematik Diplom, ähnlich der im Lehramt Mathematik. Im Gegensatz zu dort ist die höhere Spannweite allerdings nicht auf eine „Ausreißer"-Hochschule zurückzuführen, die Hochschulen im sample verteilen sich relativ gleichmäßig über den entsprechenden Bereich der Notenskala. Mit Karlsruhe und ab Mitte der 1980er Jahre auch Tübingen am oberen Ende sowie Göttingen am unteren Ende lassen sich wie im Diplom Mathematik Hochschulen identifizieren, die die Begrenzung dieser Bandbreite darstellen. Es lässt sich in den Grafiken bereits erkennen, dass sie je nach betrachtetem Zeitpunkt zwischen ca. 0.2 und 0.4 Noten über bzw. unter dem Durchschnitt aller Prüflinge liegen (Abb.90 und 91). Münster liegt über die meiste Zeit etwa in der Mitte dieser Extrempole. Die Bandbreite selbst bleibt mit ca. 0.6 Noten relativ stabil, allerdings nehmen die Minima und Maxima im Zeitverlauf zunächst ab und dann wieder zu.

Abbildung 90: Durchschnittliche Abschlussnoten an den Hochschulen in Chemie Diplom im Zeitverlauf (LOWESS 0.3)[73]



[73] In Braunschweig sind zu Beginn der Zeitreihe zwei Datenpunkte mit geringer Fallzahl (n=1) entfernt worden

Abbildung 91: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Im Mittel über den gesamten Zeitraum sind die Noten der Göttinger Prüflinge durchschnittlich 0.26 Noten besser als die aller Prüflinge, in Karlsruhe sind sie durchschnittlich 0.31 Noten schlechter. Im Zeitraum, für den das sample komplett ist (1972-2010) liegen die Göttinger Noten 0.29 Noten unter dem Durchschnitt, die Karlsruher 0.27 Noten darüber. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1960 gemittelt bei R=0.72 bzw. bei R=0.71 zwischen 1972-2010, was auch hier etwas über dem Wert liegt, der sich aus den geglätteten Daten ablesen lässt und auch gegenüber der Spannweite im Lehramt Mathematik eine Steigerung bedeutet. Für beide Zeiträume ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen 1974 (R=0.35) am geringsten und 2001 (R=1.14) am größten. Trotz der gelegentlich auftretenden Ausreißer nach oben wie unten verläuft die Spannweite in den Noten im Zeitverlauf ziemlich konstant.

In Göttingen und Braunschweig werden in der Regel die besten Noten vergeben, in Karlsruhe und Tübingen die schlechtesten. Über fünf Jahre gemittelt gibt es seit 1970 nur im letzten Abschnitt 2006-2010 in Heidelberg bessere Noten als in Göttingen oder Braunschweig und auch nur für diesen Zeitraum gibt es an einer anderen Hochschule als Karlsruhe oder Tübingen die schlechtesten Noten, und zwar in Berlin. Ansonsten liegen die Niveaus in Berlin und Heidelberg ebenso wie in Münster zwischen 1970 und 2010 gelegentlich über und gelegentlich unter denen der jeweils anderen beiden, aber immer zwischen den beiden Hochschulen, die gerade die Ober- bzw. Untergrenze bilden.

Tabelle 26: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	HD	1.62	HD	1.57	BS	1.48	GÖ	1.43	BS	1.38	GÖ	1.31	GÖ	1.26	GÖ	1.22	GÖ	1.25	HD	1.21
2	BER	1.65	GÖ	1.57	GÖ	1.51	BS	1.54	GÖ	1.42	BS	1.59	BS	1.51	BS	1.32	MS	1.34	GÖ	1.28
3	GÖ	1.75	BER	1.64	BER	1.52	BER	1.54	BER	1.68	MS	1.68	MS	1.53	MS	1.38	HD	1.40	BS	1.42
4	MS	2.00	TÜ	1.75	HD	1.55	HD	1.68	HD	1.70	BER	1.73	BER	1.62	BER	1.60	BS	1.50	MS	1.48
5	KA	2.26	MS	2.01	TÜ	1.65	TÜ	1.75	MS	1.85	HD	1.80	HD	1.78	TÜ	1.72	BER	1.51	KA	1.52
6			KA	2.13	MS	1.86	MS	1.91	TÜ	2.01	KA	1.89	KA	1.91	HD	1.76	KA	1.74	TÜ	1.81
7					KA	2.18	KA	2.11	KA	2.09	TÜ	1.91	TÜ	1.92	KA	1.86	TÜ	1.92	BER	1.92

Mit der höheren Bandbreite an Durchschnittsnoten (und höheren Fallzahlen) einhergehend, zeigen sich in der ANOVA, wiederum durchgeführt für jedes einzelne Jahr seit 1960, auch wesentlich mehr signifikante Unterschiede: In 49 der 51 getesteten Jahre unterscheiden sich mindestens zwei Hochschulen in ihrer durchschnittlichen Abschlussnote signifikant. Da auch hier für einige Jahre Varianzhomogenität vorliegt, für andere nicht, wurden die Ergebnisse der ANOVA wieder bis 1997 durch den Kruskal-Wallis Test überprüft und wieder bestätigt, dieses Mal stimmen die Testergebnisse in jedem einzelnen Jahr überein. Der Games-Howell Test, wiederum als nichtparametrische Alternative zum Kruskal-Wallis Test für die Daten ab 1998 berechnet, gibt zwei Jahre weniger als in keinem einzigen Paarvergleich signifikant different aus, womit auch unter der Anwendung nichtparametrischer Tests immer noch 47 der 51 Jahre signifikante Unterschiede im Notenniveau aufweisen.

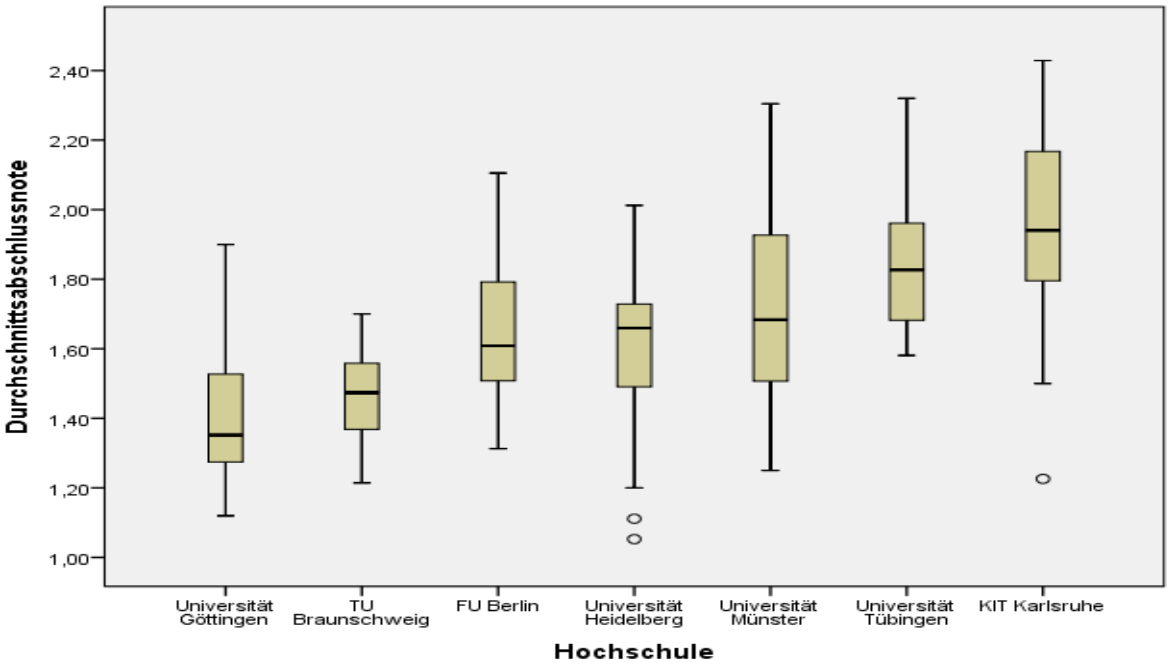
Die Noten in Göttingen und Karlsruhe unterscheiden sich am häufigsten signifikant voneinander - in 35 von 51 Jahren (bei durchschnittlich 43 bzw. 45 Prüflingen pro Jahr im Vergleichszeitraum). Berlin und Tübingen weisen die geringste Anzahl signifikanter Differenzen auf (vier von 41 Jahren bei n=29 bzw. n=38 pro Jahr). Es lässt sich, wie zuvor für die Studiengänge, die zeitliche Stabilität der signifikanten Differenzen zwischen den einzelnen Hochschulen vergleichen. Dazu wird das gleiche Verfahren angewandt wie im vorangegangen Kapitel ausführlich beschrieben: Zuerst wird für jeden Paarvergleich jede durchgängig signifikante Periode (hier: 23 Perioden) in Relation zur maximalen Dauer des Vergleichs gesetzt, bevor die Paarvergleiche anhand dieses relationalen Werts in fünf gleich breite Klassen eingeteilt werden.

Es zeigt sich dabei, dass zwischen keiner der Hochschulen eine langfristige stabile Differenz nach der hier verwendeten Definition besteht. Das stabilste Verhältnis besteht erwartungsgemäß zwischen Göttingen und Karlsruhe, dem einzigen Paarvergleich in der Klasse „längerfristig stabile Unterschiede“. Während zwischen Göttingen und Tübingen zudem noch eine mittelfristig stabile Differenz besteht, weisen ansonsten sechs der klassierten Perioden eine kurzfristige, die restlichen 15 überhaupt keine zeitliche Stabilität auf (siehe Anhang: Abb.A23).

Die Distanz zwischen den einzelnen Notenniveaus ist in Chemie deutlich breiter gefächert als in Mathematik Diplom: Zwischen Göttingen und Braunschweig ist die Distanz im Notenniveau mit 0.17 Noten Abstand im Durchschnitt über alle Jahre am geringsten, zwischen Göttingen und Karlsruhe mit einem Abstand von 0.57 Noten am größten. 14 der 21 Paarvergleiche liegen im Bereich von 0.22 bis 0.32 Noten Distanz, was dem Bereich der meisten Vergleiche in Mathematik Lehramt entspricht. Das Notenniveau in Göttingen erreicht durchschnittlich 71.9% des Karlsruher Niveaus (70.9% für den Zeitraum mit allen Hochschulen im sample), was die größte Differenz zwischen den Notenniveaus zweier Hochschulen darstellt. Die Boxplots verdeutlichen, dass sich das Notenniveau keineswegs konstant im Zeitverlauf bewegt - die Spannweite der Noten ist an allen Hochschulen mit Ausnahme der TU Braunschweig weit größer als eine halbe Note. In Münster wird sogar etwa eine ganze Note

Spannweite erreicht, und das ohne Ausreißer. Trotz der großen Spannweite an den Hochschulen wird zudem deutlich, dass in Karlsruhe und Tübingen zum größten Teil ein viel höherer Bereich im Notenspektrum abgedeckt wird als in Göttingen und v.a. in Braunschweig.

Abbildung 92: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1960-2010



Die über die Jahre gemittelten Standardabweichungen und deren Standardabweichungen offenbaren eine größere Spannweite der gemittelten Standardabweichungen als in Mathematik. Auch streuen die Noten über die Zeit etwas homogener, am stabilsten in Braunschweig, während in Karlsruhe die größte Varianz im Zeitverlauf auftritt (Spalte 3). Die Stärke der durchschnittlichen Streuung ist in Göttingen und Braunschweig deutlich geringer, in Karlsruhe etwas stärker als an den anderen Hochschulen. Insgesamt fällt sie aber etwas niedriger aus als in beiden Mathematik Studiengängen. Zwischen den beiden Vergleichszeiträumen treten keine nennenswerten Differenzen auf.

Tabelle 27: Streuung der Noten an den Hochschulen im Diplomstudiengang Chemie

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1960-2010	1972-2010	1960-2010	1972-2010
Göttingen	0.51	0.48	0.10	0.08
Braunschweig (ab 1972)	0.53	0.53	0.06	0.06
Heidelberg	0.61	0.59	0.11	0.11
Berlin	0.65	0.64	0.14	0.15
Münster	0.66	0.62	0.11	0.09
Tübingen (ab 1970)	0.69	0.69	0.08	0.08
Karlsruhe	0.76	0.71	0.15	0.12

Im Verlauf seit 1960 ist in Karlsruhe, Münster und Göttingen ein deutlicher Abwärtstrend zu erkennen. In Karlsruhe existieren zudem vergleichsweise kurze, in Münster vergleichsweise lange zyklische Bewegungen, während der Verlauf in Göttingen relativ schwankungsfrei zu verlaufen scheint. Im Vergleich der 10-Jahresmittel 2001 bis 2010 mit 1960 bis 1969 ergibt sich für Münster eine Verbesserung von -0.61 Noten, für Karlsruhe von -0.56 und für Göttingen von -0.42 Noten. In Heidelberg lässt sich eine große zyklische Bewegung mit ebenfalls verbessertem Niveau in den 2000ern gegenüber

den 1960er Jahren ausmachen (-0.31), in Braunschweig (-0.04) sind kürzere Zyklen ohne Verbesserung des Niveaus (hier gegenüber den 1970ern) sichtbar. In Berlin (+0.07) und Tübingen (+0.18) ist ebenfalls eine zyklische Struktur zu erkennen - inklusive Aufwärtstrend.

Wie auch im Vergleich der Niveaus lässt sich damit auch im Vergleich des Verlaufs eine wesentlich heterogenere Hochschullandschaft finden als in den beiden Mathematikstudiengängen. Innerhalb der Gruppen zeigt sich jedoch eine relativ einheitliche Stärke der Veränderung des Notenniveaus: An den Universitäten mit langfristiger Verbesserung beträgt sie zwischen -0.011 und -0.023 Noten pro Jahr über den gesamten Zeitraum. An den drei Hochschulen ohne Verbesserung liegt dieser Wert bei +0.002 bis +0.007 Noten. Was die kurzzeitige Stärke der Bewegungen angeht, besteht allerdings auch hier Differenzierungsbedarf: Während die Verbesserung in Münster, Karlsruhe und Göttingen recht gleichmäßig verläuft, was sich auch in einem etwa gleichzeitig beginnenden und endenden stärkstem Trendbereich äußert, ist die Abwärtsdynamik in Heidelberg in dem vergleichsweise kurzen Zeitraum von 1990-2007 doppelt bis dreimal so stark wie an den anderen drei Standorten. Der Aufwärtstrend in Tübingen ist von einer kurzen starken Phase zwischen 1974-1982 dominiert, die im Durchschnitt die zweieinhalbfache Stärke der dynamischsten Berliner Aufwärtsbewegung erreicht.

Die Trendbereinigung (Abb.93) macht sichtbar, dass auch der relativ glatte Verlauf in Göttingen zyklische Bewegungen enthält, wenn auch wesentlich schwächer ausgeprägt als bei den anderen Hochschulen. Dass die Göttinger Reihe in der Grafik nicht gleichmäßig um den Nullpunkt verläuft, sondern etwas darunter, liegt daran, dass die Originalreihen, die zur Berechnung verwendet wurden, länger zurückreichen als bis zum Beginn der hier erfolgten Darstellung 1960 und die Werte der trendbereinigten Reihe vor 1960 entsprechend häufiger über dem Nullpunkt liegen[74]. Auch in Heidelberg verlaufen noch Zyklen um die große zyklische Abwärtsbewegung herum. Die großen Zyklen weisen auch hier eine ungefähre Länge von ca. 20 Jahren auf, in Braunschweig und Tübingen treten zudem noch kürzere Schwankungen auf.

Tabelle 28: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Chemie 1960-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Münster	-0.61	-0.64	-0.013	-0.94 (1960-2004)	-0.021
Karlsruhe	-0.56	-1.15	-0.023	-1.20 (1966-2010)	-0.027
Göttingen	-0.42	-0.54	-0.011	-0.78 (1962-2005)	-0.018
Heidelberg	-0.31	-0.63	-0.013	-0.96 (1990-2007)	-0.056
Braunschweig (ab 1972)	-0.04	+0.07	+0.002	--	--
Berlin	+0.07	+0.36	+0.007	+0.73 (1980-2007)	+0.027
Tübingen (ab 1970)	+0.18	+0.17	+0.004	+0.58 (1974-1982)	+0.073

[74] Es gilt auch für alle folgenden Abbildungen trendbereinigter Zeitreihen, die nicht gleichmäßig um den Nullpunkt verlaufen, dass sich ein gleichmäßiges Bild ergibt, wenn die Abbildung um die davor liegenden Jahre erweitert wird.

Tabelle 29: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Chemie 1972-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Karlsruhe	-0.50	-1.18	-0.031	-1.18 (1972-2010)	-0.031
Münster	-0.49	-0.41	-0.011	-0.86 (1977-2004)	-0.032
Heidelberg	-0.34	-0.30	-0.008	-0.96 (1990-2007)	-0.056
Göttingen	-0.18	-0.30	-0.008	-0.51 (1973-2005)	-0.016
Braunschweig	-0.04	+0.07	+0.002	--	--
Berlin	+0.21	+0.56	+0.015	+0.73 (1980-2007)	+0.027
Tübingen	+0.12	+0.23	+0.006	+0.58 (1974-1982)	+0.073

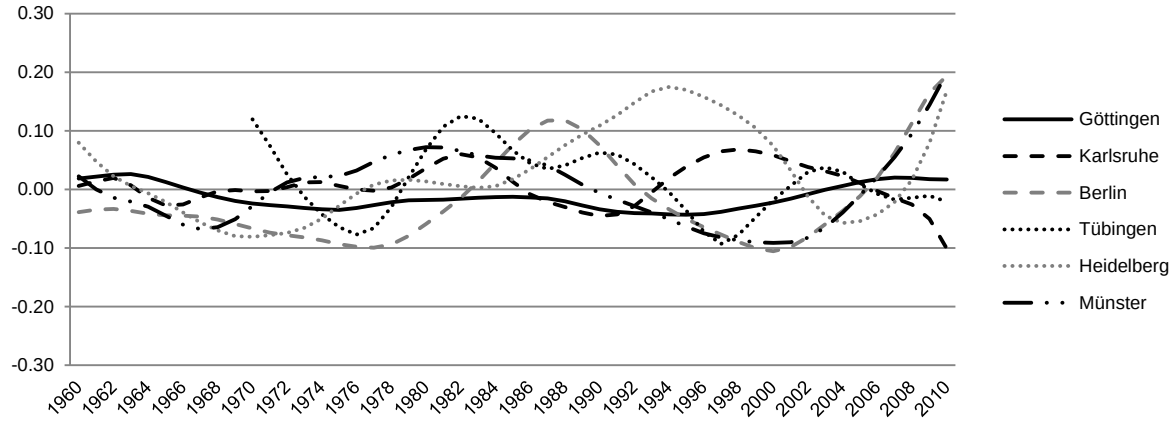
Die Betrachtung der Notenentwicklung an den einzelnen Hochschulen entspricht somit im Gegensatz zu Mathematik nicht unbedingt dem Verlauf der Noten auf Studiengangebene. Nur an vier der sieben Universitäten (Göttingen, Karlsruhe, Münster, Heidelberg) ist ein deutlicher Abwärtstrend, begleitet von schwächer oder stärker ausgeprägten Wellen, zu beobachten. Göttingen, Karlsruhe und Münster prägen den über alle Prüflinge gemittelten Verlauf der leicht zyklischen Abwärtsbewegung. Während die Braunschweiger Zyklenbewegung auf konstantem Niveau und die Heidelberger Verbesserung mit langer, zyklisch umlagerter Abwärtskurve noch als leichte Abweichung vom Studiengangverlauf gesehen werden können, passen die Verschlechterungen in Berlin und Tübingen nicht ins Studiengangmuster.

Dass deren konträre Verläufe im Studiengangmittel nicht sichtbar werden, ist mit der vergleichsweise geringen Fallzahl im Gegensatz zu den vier Universitäten mit Abwärtstrend erklärbar. Letztere weisen mit durchschnittlich 214 Prüflingen pro Jahr seit 1960 im Vergleich zu Berlin und Tübingen mit zusammen durchschnittlich 69 Prüflingen etwa dreimal so viel Gewicht in der jährlichen Mittelwertbildung auf. In Heidelberg ist von 1975 bis 1988 eine Plateauphase vorhanden, die zeitlich mit dem Plateau auf Studiengangebene übereinstimmt. Auffällig ist, dass solche Phasen mit maximalen Schwankungen im Bereich von 0.20 Noten sich überschneidend in Tübingen (1987-1994), Karlsruhe (1987-1999), Göttingen (1986-2004) und Münster (1997-2007) existieren, während sich die Noten auf Studiengangebene in dieser Zeit verbessern.

Diese anhaltende Verbesserung beruht fast ausschließlich auf der Verbesserung in Heidelberg, welche von der Mitte der 1980er Jahre an ihre stärkste Phase hat. Die Noten dort verlaufen im Zeitverlauf in einer großen zyklischen Bewegung, die unter dem Durchschnitt aller Prüflinge beginnt und diesen dann zunächst in der Aufwärts-, dann in der Abwärtsbewegung schneidet. Karlsruhe als Extrem am oberen Ende passt sich im Zeitverlauf langsam an den Durchschnitt aller Prüflinge an und wird schließlich von Tübingen, dass den entgegengesetzten Verlauf von der Mitte zum oberen Ende der Bandbreite nimmt, als Hochschule mit den schlechtesten Noten abgelöst. In Münster verlaufen die Noten relativ mittig zwischen Karlsruhe und Göttingen, wobei es im Zeitverlauf von einer leichteren Abweichung Richtung Karlsruhe zu einer Annäherung an Göttingen kommt. Die Noten dort markieren konstant das untere Ende der Bandbreite und passen sich in einer leichten Kurvenbewegung mit konstantem Abstand dem Absinken und Ansteigen der Maximalwerte beim Übergang von Karls-

ruhe zu Tübingen am oberen Ende an. Das Braunschweiger Notenniveau gleicht sich im Zeitverlauf dem Gesamtdurchschnitt an, wobei es genau genommen umgekehrt ist: Das sinkende Gesamtmittel nähert sich zunehmend dem recht konstanten Niveau in Braunschweig an. Insgesamt ist das Verhältnis der Noten zwischen den einzelnen Hochschulen damit dynamischer und vor allem ungleichmäßiger als in Mathematik.

Abbildung 93: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.3-LOWESS 0.9)



Biologie Diplom[75]

Die Spannweite der Durchschnittsnoten ist, wie schon die grafische Darstellung in Biologie erkennen lässt, kleiner als in Mathematik und Chemie. Bis Mitte der 1990er Jahre lässt sich zwischen unterem und oberem Ende ein Abstand von ca. 0.2 bis 0.4 Noten ablesen, mit einem leichten Trend zur Verringerung. Anschließend brechen Karlsruhe nach oben sowie Braunschweig und Heidelberg nach unten kurzfristig aus, wodurch sich die Bandbreite zwischenzeitlich auf ca. 0.6 Noten Abstand erhöht (Abb.94 und 95). Es ist allerdings keine Hochschule zu erkennen, die außerhalb dieses kurzen Zeitraums alleine die obere Begrenzung der Bandbreite darstellt, während am unteren Ende ca. 20 Jahre lang Berlin den Extrempol einnimmt, bevor das Niveau mehrerer Hochschulen unter das Berliner Niveau sinkt. Die grafische Darstellung legt nahe, dass die Universitäten näher aneinander liegen als in Mathematik und Chemie.

[75] Anmerkungen zur Datenbasis: Da der Diplomabschluss in Biologie an vielen Hochschulen in Deutschland erst in den 1960er Jahren eingeführt wurden und In Göttingen, Braunschweig, Karlsruhe und Heidelberg zu Beginn der Zeitreihen je ein, in Berlin zwei Datenpunkte mit geringer Fallzahl (n<4) entfernt worden sind, stehen Daten für mindestens zwei Hochschulen erst ab 1969 zur Verfügung. Für Heidelberg fehlt in der Prüfungsstatistik der Wert des Jahres 2010, in Münster die Werte der Jahre 2009 und 2010.

Abbildung 94: Durchschnittliche Abschlussnoten an den Hochschulen in Biologie Diplom - Zeitverlauf (LOWESS 0.3)

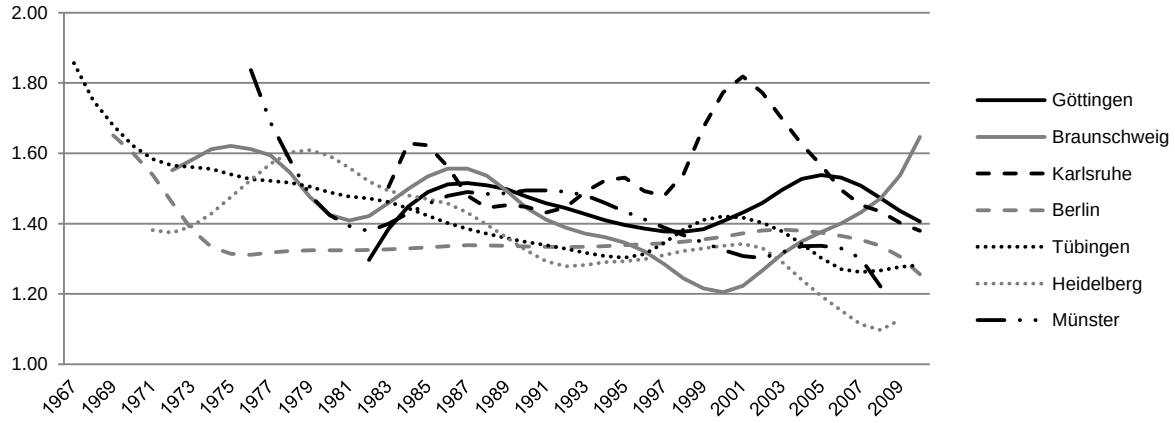
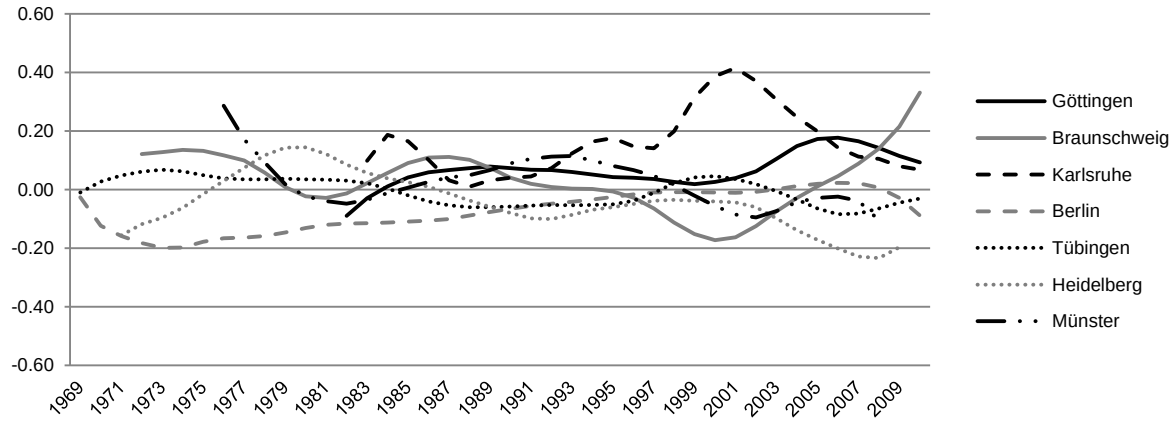


Abbildung 95: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Die Originaldaten zeigen, dass Karlsruhe mit durchschnittlich 0.16 Noten über dem Durchschnitt aller Prüflinge liegt (0.17 Noten für den Zeitraum von 1983 bis 2008, für den alle Hochschulen im sample sind), Berlin mit 0.08 (0.03) Noten darunter. Im Zeitraum von 1983 bis 2008 liegen die Noten in Heidelberg mit 0.08 unter dem Studiengangmittel am weitesten unter dem Durchschnitt.

Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1960 gemittelt bei R=0.42 bzw. bei R=0.39 zwischen 1983-2008, was ungefähr dem Wert entspricht, der sich aus den geglätteten Daten ablesen lässt, womit sie wie erwartet die Werte für Mathematik und Chemie unterbietet. Für beide Zeiträume ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen 1995 (R=0.15) am geringsten und 2000 (R=0.91) am größten. Diese Extrempunkte entsprechen dem Gesamtverlauf der Spannweite, die zunächst bis in die 1990er abnimmt und dann sprunghaft ansteigt.

Die 5-Jahresmittel der Durchschnittsnoten bestätigen den bisherigen Eindruck: In Berlin werden bis zur zweiten Hälfte der 1980er Jahre die besten Noten vergeben, danach übernimmt diese Rolle mit Ausnahme des Zeitraums 1996-2000 Heidelberg. Die schlechtesten Noten gibt es ab 1996 entweder in Karlsruhe oder in Göttingen. Vorher ist kein Muster zu erkennen.

Tabelle 30: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

	1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	BER	1.42	BER	1.30	BER	1.29	BER	1.34	BER	1.35	HD	1.28	BS	1.25	HD	1.31	HD	1.08
2	TÜ	1.51	HD	1.38	TÜ	1.50	MS	1.40	TÜ	1.36	TÜ	1.32	HD	1.32	BS	1.32	TÜ	1.27
3			TÜ	1.56	BS	1.56	BS	1.42	HD	1.41	BER	1.32	TÜ	1.34	MS	1.32	MS	1.31
4			BS	1.60	MS	1.58	GÖ	1.43	KA	1.46	BS	1.35	BER	1.36	BER	1.37	BER	1.35
5					HD	1.62	HD	1.48	MS	1.47	GÖ	1.42	MS	1.36	TÜ	1.40	KA	1.43
6							TÜ	1.49	GÖ	1.51	KA	1.47	GÖ	1.38	GÖ	1.50	BS	1.46
7							KA	1.66	BS	1.55	MS	1.48	KA	1.61	KA	1.68	GÖ	1.47

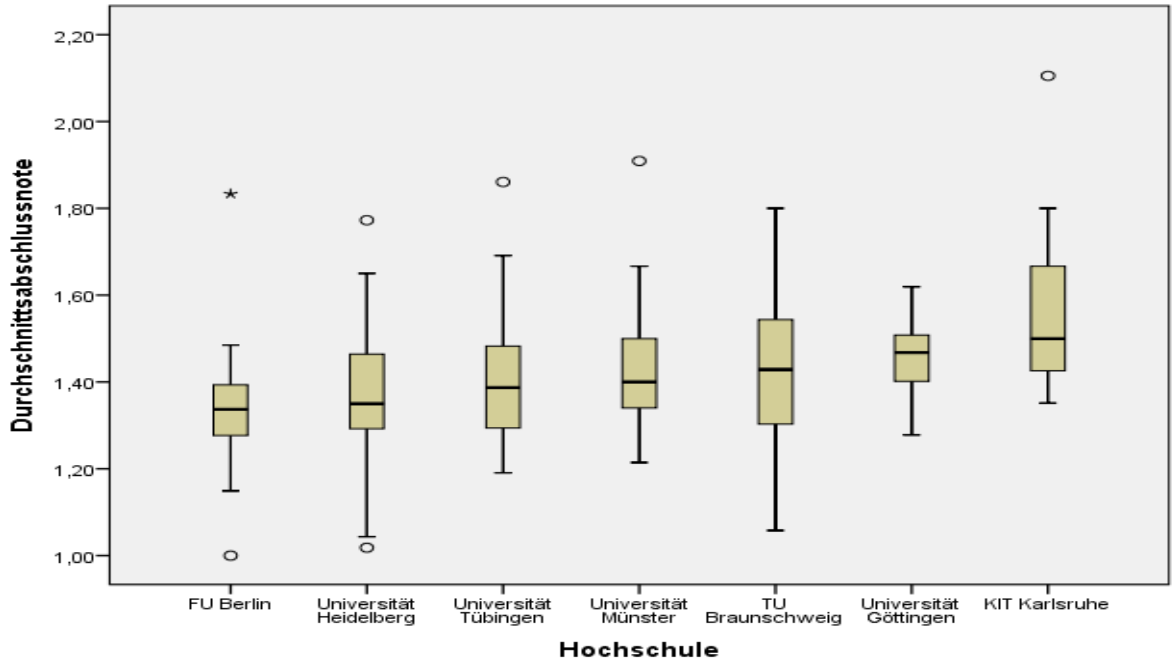
Mit 22 Jahren (erst seit 1969) liegt die Anzahl der Jahre, in denen sich anhand der Varianzanalyse (T-Tests für die Jahre 1969 und 1970) signifikante Unterschiede in den Durchschnittsnoten nachweisen lassen, zwischen denen in Mathematik und Chemie. Anhand des Kruskal-Wallis bzw. des Games-Howell Tests ergeben sich 28 Jahre mit signifikanten Unterschieden, was erstmals eine deutliche Abweichung zwischen den Tests bedeutet. In den sechs Jahren, in denen die Testergebnisse abweichen, liegt dreimal Varianzhomogenität vor, dreimal nicht. Entsprechend scheint es sinnvoll, in der einen Hälfte der zweifelhaften Fälle auf die Ergebnisse der ANOVA in der anderen Hälfte auf die des Kruskal-Wallis Test zu vertrauen. Dadurch ergibt sich eine Gesamtzahl von 25 Jahren mit signifikanten Unterschieden zwischen den Hochschulen. Die Noten in Heidelberg und Karlsruhe unterscheiden sich am häufigsten signifikant voneinander - in neun von 27 Jahren (bei durchschnittlich 96 bzw. 32 Prüflingen pro Jahr im Vergleichszeitraum). Göttingen und Heidelberg (neun von 28 Jahren) sowie Göttingen und Tübingen (neun von 29 Jahren) weisen ein ähnliches Verhältnis auf.

Ein Vergleich der zeitlichen Stabilität der signifikanten Differenzen zwischen den einzelnen Hochschulen offenbart, dass zwar 11 der insgesamt 21 Paarvergleiche nach der hier verwendeten Definition durchgängig stabile Perioden in den signifikanten Notenniveaudifferenzen umfassen, welche allerdings alle nur von äußerst geringer Dauer sind: Selbst kurzfristig stabil sind gerade einmal die Differenzen zweier Paarvergleiche (Göttingen/Tübingen und Karlsruhe/Heidelberg) - die anderen neun durchgängigen Perioden sind allesamt so kurz, dass sie als nicht zeitlich stabil einzustufen sind (siehe Anhang: Abb.A24).

Die Distanz zwischen den einzelnen Notenniveaus ist in Biologie deutlich geringer als in Mathematik und Chemie: Zwischen Göttingen und Münster ist die Distanz im Notenniveau mit 0.10 Noten Abstand im Durchschnitt über alle Jahre am geringsten, zwischen Heidelberg und Karlsruhe mit einem Abstand von 0.26 Noten am größten. Alle anderen 19 Paarvergleiche liegen im Bereich von durchschnittlich 0.12 bis 0.22 Noten Distanz. Das Notenniveau in Heidelberg erreicht durchschnittlich 84.8% des Karlsruher Niveaus (85.1% für den Zeitraum mit allen Hochschulen im sample), was die größte Differenz zwischen den Notenniveaus zweier Hochschulen darstellt.

Die hohe Homogenität der einzelnen Hochschulen hinsichtlich der vergebenen Abschlussnoten äußert sich auch in ihrer Verteilung über die Zeit. Alle Hochschulen decken in etwa das gleiche Notenspektrum ab. Karlsruhe hebt sich etwas zu den schlechteren Noten hin ab und in Heidelberg und Braunschweig reicht die Verteilung im Zeitverlauf etwas weiter in Richtung des 1.00er Bereichs als an den anderen Hochschulen, wobei in Braunschweig auch in Richtung schlechterer Durchschnitte ein etwas größerer Bereich in Anspruch genommen wird - dafür sind allerdings dort keine Ausreißer zu verzeichnen. In Berlin und Göttingen ist der Bereich, in dem sich die Notendurchschnitte bewegen am geringsten.

Abbildung 96: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1969-2010



Auch die Spannweite der gemittelten Standardabweichungen ist sehr homogen und die mittlere Streuung äußerst gering. Es gibt keine klaren Ausreißer und selbst der höchste Wert von s=0.57 in Karlsruhe liegt unter jedem Wert der beiden Mathematikstudiengänge und wird nur von zwei Werten in Chemie unterboten. Über die Zeit streuen die Noten auch recht schwach, aber unterschiedlich stabil, am einheitlichsten in Göttingen und am weitesten in Heidelberg - etwa dreimal so stark.

Tabelle 31: Streuung der Noten an den Hochschulen im Diplomstudiengang Biologie

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1969-2010	1983-2008	1969-2010	1983-2008
Berlin	0.52	0.55	0.10	0.06
Tübingen	0.53	0.49	0.08	0.05
Heidelberg (1971-2009)	0.53	0.49	0.13	0.13
Göttingen (ab 1982)	0.55	0.55	0.04	0.04
Münster (1976-2008)	0.55	0.54	0.08	0.07
Braunschweig (ab 1972)	0.55	0.53	0.11	0.11
Karlsruhe (ab 1983)	0.57	0.57	0.09	0.09

Wird die Entwicklung der Noten an den einzelnen Hochschulen seit 1960 im Verlauf betrachtet, so zeigt sich in Tübingen, Heidelberg und Münster ein durchgängiger Abwärtstrend, in Heidelberg und

Münster mit offensichtlicher zyklischer Komponente. Auch in Braunschweig ist ein zyklisch geprägter Abwärtstrend zu beobachten, dieser endet allerdings bereits Anfang der 2000er und zieht einen deutlichen Anstieg nach sich, der schließlich wieder ungefähr bis zum Ausgangsniveau 1972 zurückführt. In Berlin ist nur in den wenigen Jahren zu Beginn der Reihe ein starkes Absinken des Notenniveaus zu beobachten, ab Anfang der 1970er jedoch ist das Niveau dort weitestgehend stabil und steigt eher noch minimal an, als dass es sinkt. In Göttingen und Karlsruhe ist schließlich ein zyklischer Verlauf ohne langfristige Verbesserung festzustellen.

Im Vergleich des Mittelwerts der letzten 10 Jahre mit dem der ersten 10 der jeweiligen Reihe ergibt sich für Münster eine Verbesserung von -0.18 Noten und eine durchschnittliche jährliche Veränderung von -0.022 Noten - die auch der maximalen Trendstärke entspricht, da der Trend vom ersten (1976) bis zum letzten Zeitpunkt (2008) durchgängig vorliegt. Tübingen und Heidelberg weisen eine höhere Verbesserung im 10-Jahresvergleich auf, bei einer ähnlichen maximalen Abwärtsdynamik. Die ist auch in Braunschweig in der maximalen Verbesserungsperiode auf einem ähnlichen Level. Hier zeigt sich, dass das mittlere Niveau der letzten 10 Jahre der Reihe trotz Aufwärtstrend in den 2000ern unter dem Niveau der ersten 10 Jahre in den 1970ern liegt - auch deshalb, weil in der Zeit, in der das Notenniveau sinkt, die höchste Trendstärke vorliegt - bei allen vier Hochschulen mit Abwärtstrend etwa zeitgleich. In Berlin zeigt sich nach dem kurzen Absturz der Noten Ende der 1960er Jahre auch in den Zahlen ein leichter Aufwärtstrend, in Göttingen und Karlsruhe gleicht der zyklische Verlauf das Niveau im Zeitverlauf weitestgehend aus. Die Trendbereinigung (Abb.97) offenbart auch hier 10-20 jährige Zyklen, in Göttingen und Karlsruhe tendenziell kürzer als in Tübingen, Münster, Heidelberg und Braunschweig, an den letzten beiden Hochschulen ausgeprägter als an den übrigen.

Tabelle 32: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Biologie 1969-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Heidelberg (1971-2009)	-0.28	-0.20	-0.005	-0.75 (1978-2007)	-0.026
Tübingen	-0.20	-0.17	-0.004	-0.50 (1976-2003)	-0.019
Münster (1976-2008)	-0.18	-0.70	-0.022	-0.70 (1976-2008)	-0.022
Braunschweig (ab 1972)	-0.16	-0.10	-0.003	-0.74 (1975-2001)	-0.029
Göttingen (ab 1982)	± 0.00	+0.14	+0.005	--	--
Berlin	+0.03	-0.59	-0.015	+0.48 (1970-2003)	+0.015
Karlsruhe (ab 1983)	+0.06	-0.15	-0.006	--	--

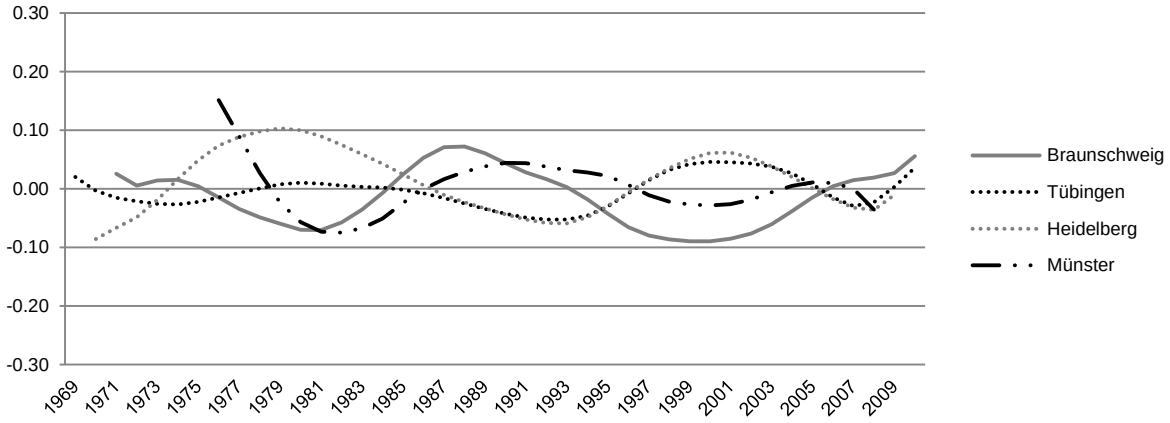
Tabelle 33: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Biologie 1983-2008

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Braunschweig	-0.20	-0.09	-0.004	-0.68 (1986-2001)	-0.045
Münster	-0.17	-0.17	-0.007	-0.36 (1989-2008)	-0.019
Heidelberg	-0.13	-0.32	-0.012	-0.53 (1987-2007)	-0.027
Tübingen	-0.03	-0.17	-0.007	-0.53 (1985-2003)	-0.018
Göttingen	-0.01	+0.05	+0.002	--	--
Berlin	+0.06	+0.12	+0.005	+0.24 (1983-2003)	+0.012
Karlsruhe	+0.13	-0.14	-0.006	--	--

Die Notenentwicklung an den einzelnen Hochschulen entspricht damit auch in Biologie nur teilweise dem Verlauf der Noten auf Studiengangebene. Dem Abwärtstrend an vier Hochschulen, begleitet von schwächer oder stärker ausgeprägten Wellen, stehen drei Hochschulen ohne durchgängige Verbesserung entgegen. Dass der Abwärtstrend in Braunschweig gegen Ende der Reihe in einen Aufwärtstrend umschlägt begünstigt die durch die Verläufe in Göttingen, Karlsruhe und besonders Berlin entstehende Stabilität des Notenniveaus ab Mitte der 1970er auf Studiengangebene.

Interessant ist, dass im Gegensatz zu den bisher betrachteten Studiengängen, in denen die Spannweite der Notenniveaus über den Zeitverlauf relativ stabil ist, in Biologie Mitte der 1990er Jahre eine Ausdifferenzierung der Notenniveaus auf der Notenskala zu beobachten ist, verursacht durch die Verschlechterung der Noten in Karlsruhe und durch die Verbesserungen in Heidelberg und Braunschweig im letzten Abschnitt des beobachteten Zeitraums.

Abbildung 97: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.4-LOWESS 0.9)



Psychologie Diplom

In Psychologie zeigt sich im Zeitverlauf eine dynamische Spannweite im Notenniveau der einzelnen Hochschulen. In den 1960er Jahren noch bis ca. 0.7 Noten Unterschied reichend, geht dieser Wert in einem ähnlichen Verlauf wie in Biologie bis in die 1990er zunächst bis auf ca. 0.2 zurück, bevor sich der maximale Abstand wieder erhöht und in den 2000ern etwas unter 0.5 Noten Differenz liegt. Die zunächst erfolgende Annäherung der besten und schlechtesten Durchschnittsnoten kommt durch die Aufwärtsbewegung der über 20 Jahre besten Noten in Berlin ab dem Beginn der 1980er Jahre in Kombination mit dem allgemeinen Abwärtstrend an den anderen fünf Hochschulen zustande. Die sich anschließende Vergrößerung der Spannweite entsteht durch eine späte Stagnation der Noten in Tübingen bei gleichzeitig noch deutlicher Verbesserung in Heidelberg. Die untere Grenze des Notenniveaus wird in Psychologie in den 1960ern zunächst von Münster, zwischen den 1970ern und 1990ern von Berlin und in den 2000ern dann von Heidelberg eingenommen. Die Extremposition am oberen Ende der Bandbreite lässt sich anhand der grafischen Darstellung nicht eindeutig (einer) ein-

zelnen Hochschule(n) zuordnen. Braunschweig liegt über den gesamten Zeitverlauf am nächsten am Notendurchschnitt aller Prüflinge.

Abbildung 98: Durchschnittliche Abschlussnoten an den Hochschulen in Psychologie Diplom - Zeitverlauf (LOWESS 0.3)

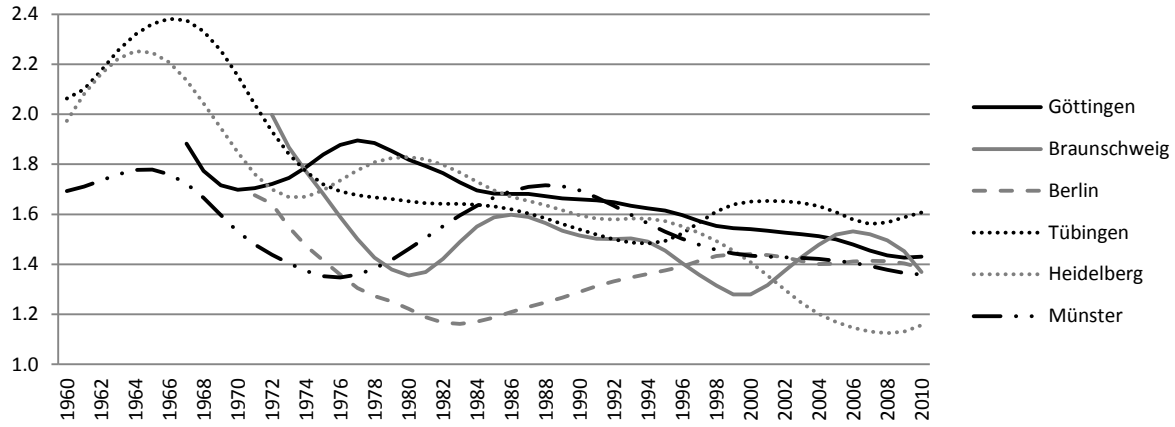
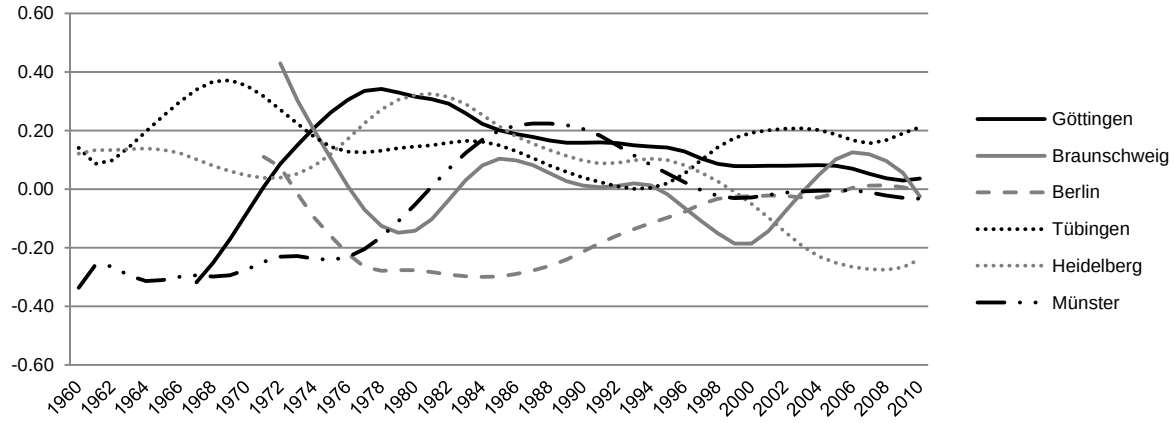


Abbildung 99: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Aufgrund der längsten Periode am unteren Ende der Notenspannweite liegen die Berliner Noten mit 0.13 Noten am weitesten unter dem Mittel aller Prüflinge, während die Tübinger Noten über den gesamten Zeitraum betrachtet 0.16 Noten schlechter als der Studiengangdurchschnitt sind. Im Zeitraum von 1972 bis 2010, in dem für alle sechs Hochschulen vollständige Daten vorliegen, liegen die Göttinger Noten im Durchschnitt am weitesten über dem Mittel aller Absolvent*innen, sie sind 0.17 Noten schlechter. Die Noten in Berlin sind auch in diesem Zeitraum die besten, sie sind 0.14 Noten besser. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1960 gemittelt bei R=0.57 bzw. bei R=0.53 zwischen 1972-2010. Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen in beiden Zeiträumen 1996 (R=0.23), über den gesamten Zeitraum betrachtet am größten 1966 (R=1.33) und im Zeitraum mit allen Hochschulen im sample 2001 (0.87). Auch hier entsprechen die Extrempunkte damit der anfänglichen Abnahme und dem anschließenden Anstieg der Spannweite im Zeitverlauf.

Über jeweils 5-Jahresabschnitte gemittelt zeigt sich keine konstante Notenhierarchie. Für den Zeitraum von 1971-1995 lässt sich sagen, dass es dort in Göttingen entweder die schlechtesten oder

zweitschlechtesten Noten gibt. In Berlin gibt es, wie auch schon aus der Grafik abzulesen, von 1976-1995 die besten Noten. Wie ebenfalls bereits in der Grafik zu erkennen, werden die schlechtesten Noten zwischen 1996 und 2010 in Tübingen vergeben.

Tabelle 34: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	MS	1.85	MS	1.64	MS	1.32	BER	1.28	BER	1.16	BER	1.25	BER	1.35	BS	1.30	HD	1.23	HD	1.11
2	TÜ	2.17	GÖ	1.70	BER	1.54	MS	1.34	BS	1.51	BS	1.55	TÜ	1.44	MS	1.41	BER	1.40	MS	1.37
3	HD	2.21	HD	2.12	HD	1.54	BS	1.44	MS	1.62	HD	1.61	BS	1.49	BER	1.44	BS	1.43	BER	1.41
4			TÜ	2.48	TÜ	1.76	TÜ	1.66	TÜ	1.62	TÜ	1.63	HD	1.57	HD	1.54	MS	1.45	GÖ	1.43
5					BS	1.78	GÖ	1.89	GÖ	1.71	GÖ	1.67	MS	1.63	GÖ	1.54	GÖ	1.55	BS	1.50
6					GÖ	1.79	HD	1.90	HD	1.76	MS	1.74	GÖ	1.64	TÜ	1.55	TÜ	1.69	TÜ	1.58

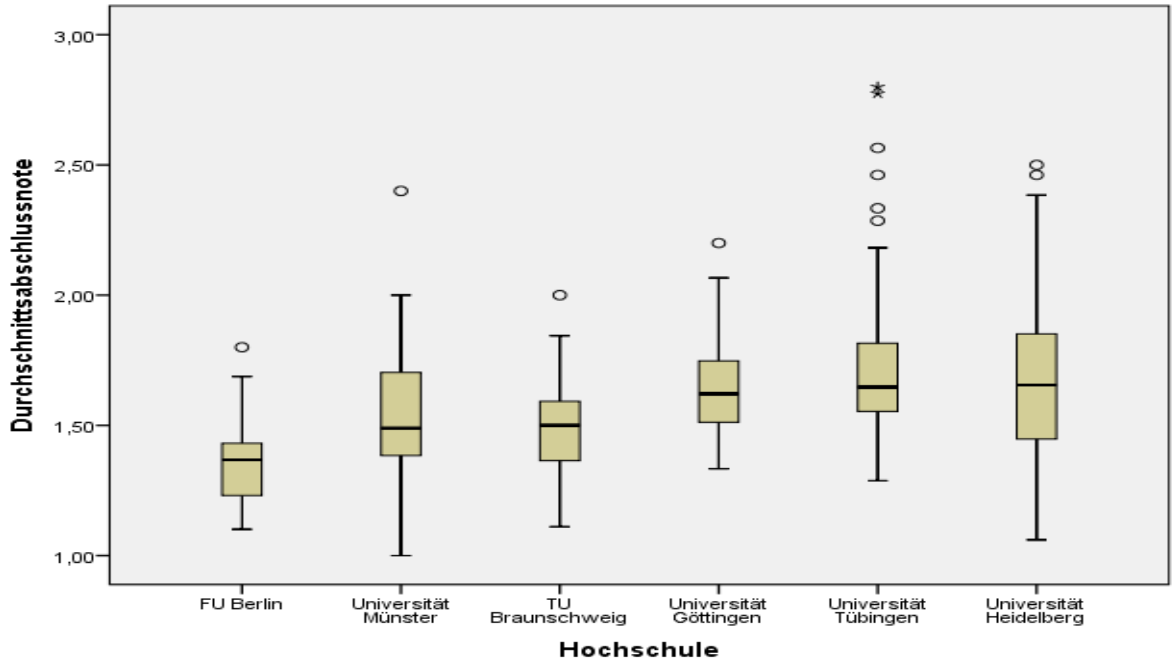
Die einfaktorielle ANOVA ergibt trotz der geringeren Spannweite im Vergleich etwa zu VWL in 45 von 51 Jahren signifikante Unterschiede in den Durchschnittsnoten zwischen mindestens zwei Hochschulen, vermutlich bedingt durch die jeweilige Ausreißerposition einer der Universitäten nach unten. Die kombinierte Überprüfung mit Kruskal-Wallis und Games-Howell Test ergibt exakt die gleichen Ergebnisse. Es unterscheiden sich die Noten in Heidelberg und Berlin am häufigsten signifikant voneinander - in 24 von 40 Jahren, bei einer durchschnittlichen Fallzahl von 74 bzw. 140 Prüflingen pro Jahr im Vergleichszeitraum.

Aus den 15 Paarvergleichen lassen sich 18 Perioden ablesen, in denen drei oder mehr aufeinanderfolgende Jahre signifikante Differenzen aufweisen. Von diesen 18 Perioden sind drei als mittelfristig stabile einzustufen – allesamt unter der Beteiligung von Berlin und zwar in den Kombinationen mit Göttingen, Tübingen und Heidelberg. Zwischen Berlin und Münster sowie zwischen Tübingen und Münster besteht zudem je eine Phase kurzfristiger Stabilität bezüglich der Differenzen im Notenniveau. Die übrigen 13 Phasen sind so kurz, dass sie sich nicht als zeitlich stabil einordnen lassen (siehe Anhang: Abb.A25).

Die Distanzen (Betragsfunktion der Differenz) zwischen den einzelnen Hochschulen reichen von 0.19 zwischen Braunschweig und Münster bis 0.35 zwischen Berlin und Heidelberg. 12 der 15 Paarvergleiche liegen im Bereich von 0.19 bis 0.28 Noten Distanz - ähnliche Werte wie in Mathematik Diplom und nach denen in Biologie mit die niedrigsten. Die größte prozentuale Differenz zwischen den Notenniveaus lässt sich über die Jahre gemittelt zwischen Berlin und Göttingen feststellen: Das Notenniveau in Berlin erreicht durchschnittlich 83.1% des Göttinger Niveaus (82.8% im Zeitraum von 1972-2010).

In Berlin, Münster, Braunschweig und Heidelberg decken die Durchschnittsnoten in ihrer Verteilung über die Zeit entweder fast oder sogar vollständig den Bereich zwischen 1.50 und 1.00 ab, in Göttingen und Tübingen endet die Spannweite innerhalb dieses Bereichs. In Heidelberg reichen die Durchschnittsnoten fast von 2.50 bis 1.00, was eine starke Dynamik im Zeitverlauf impliziert, während die Berliner Noten sich am homogensten über die Zeit verteilen.

Abbildung 100: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1960-2010



Die Standardabweichungen der gemittelten Standardabweichungen zeigen, dass die Noten über die Zeit unterschiedlich stark streuen (Spalten 4 und 5). In Heidelberg ist die Streuung der Streuung mit einem anderthalbfach bis doppelt so hohen Wert wie an vier der fünf anderen Standorte am stärksten, auch Tübingen liegt über den anderen Werten. Beide Werte reduzieren sich allerdings deutlich, wird nur den Zeitraum ab 1972 herangezogen, in dem für alle Hochschulen Daten vorliegen. Bezüglich der Stärke der durchschnittlichen Streuung heben sich Münster und Tübingen etwas von den anderen vier Hochschulen ab, die mit vergleichsweise niedrigen Werten, insgesamt etwas über dem durchschnittlichen Niveau in Biologie, nahe beieinander liegen. Auch hier passen sich die erhöhten Werte allerdings an das Niveau der anderen Hochschulen an, wenn nur der Zeitraum ab 1972 im Fokus steht - dann ist auch das etwas niedrigere Niveau im Studiengang Biologie erreicht, für den an den meisten Hochschulen ja erst später Daten vorliegen.

Tabelle 35: Streuung der Noten an den Hochschulen im Diplomstudiengang Psychologie

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1960-2010	1972-2010	1960-2010	1972-2010
Berlin (ab 1971)	0.51	0.51	0.09	0.09
Heidelberg	0.54	0.52	0.15	0.11
Braunschweig (ab 1972)	0.55	0.55	0.10	0.10
Göttingen (ab 1967)	0.56	0.56	0.08	0.06
Münster	0.60	0.57	0.10	0.06
Tübingen	0.62	0.57	0.13	0.06

Der Notenverlauf über die Zeit ist an den Hochschulen in Psychologie durch eine kurze aber starke Abwärtsdynamik bis in die 1970er Jahre geprägt. Lediglich in Heidelberg setzt sie sich über den gesamten Zeitraum fort, was zum größten Niveauunterschied zwischen 2000er und 1960er Jahren führt. In Göttingen beginnt die Abwärtsbewegung erst in den 1970ern richtig, während sie an den anderen Hochschulen dort schon die deutlichste Verbesserungsphase hinter sich hat. In Münster und

Berlin ist der stärkste Trendbereich schon 1976 bzw. 1980 beendet, in Berlin ist danach sogar eine Aufwärtsphase zu beobachten. In der Zeit, in der sich die Noten abwärts bewegen, geschieht dies in Psychologie mit einer enormen Abwärtsdynamik, in der die Noten fast bis zur unteren Begrenzung der Notenskala fallen, in Berlin mit durchschnittlich 0.1 Noten pro Jahr zwischen 1973 und 1980. Auch in Münster und, schon nur noch halb so stark wie in Berlin, in Tübingen sind noch hohe jährliche Verbesserungen im stärksten Trendbereich festzuhalten, bevor die durchschnittliche Veränderung in Braunschweig, Heidelberg und Göttingen Werte erreicht, die auch in den anderen Studiengängen auftreten. Hier wirkt die Verbesserung allerdings über einen so langen Zeitraum, dass die absoluten Veränderungen im stärksten Trendbereich weit über den Werten der meisten anderen Entwicklungen liegen.

Tabelle 36: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Psychologie 1960-2010

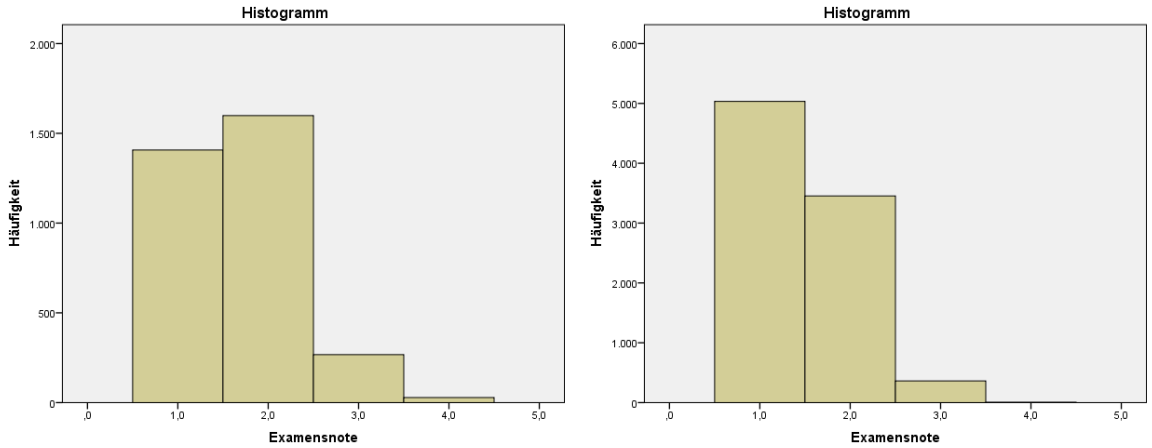
Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Heidelberg	-1.04	-0.64	-0.013	-1.44(1962-2008)	-0.031
Tübingen	-0.71	-0.69	-0.014	-1.51 (1966-1992)	-0.058
Münster	-0.36	-0.38	-0.007	-1.19 (1963-1976)	-0.091
Göttingen (ab 1967)	-0.27	-0.21	-0.005	-0.86 (1968-2008)	-0.021
Braunschweig (ab 1972)	-0.08	-0.62	-0.016	-0.89 (1972-1999)	-0.033
Berlin (ab 1971)	-0.00	-0.21	-0.005	-0.70 (1973-1980)	-0.100

Tabelle 37: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang Psychologie 1972-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Heidelberg	-0.56	-0.08	-0.002	-0.98 (1979-2008)	-0.034
Göttingen	-0.38	-0.22	-0.006	-0.73 (1979-2008)	-0.025
Braunschweig	-0.08	-0.62	-0.016	-0.89 (1972-1999)	-0.033
Tübingen	-0.06	-0.19	-0.005	-0.50 (1972-1992)	-0.025
Berlin	+0.04	-0.32	-0.008	-0.70 (1973-1980)	-0.100
Münster	+0.07	+0.01	±0.000	-0.52 (1988-1997)	-0.057

Nach dem starken Absinken bleiben die Noten in Braunschweig, Tübingen und Münster in zyklischen Bewegungen von 10-20 Jahren, die sich wiederum an allen Hochschulen außer Berlin zeigen (Abb.102), relativ stabil. In Berlin steigen sie sogar wieder leicht und erreichen dort in den 2000ern wieder das Niveau der 1970er. Dies mag daran liegen, dass die Noten mit einem Durchschnitt im Einserbereich bereits so gut sind, dass ein weiterer Abwärtstrend kaum mehr möglich ist. Ein Blick auf die Notenverteilungen vor und ab 1980 verdeutlicht dies: Von 1980 bis 1997 wurde an allen betrachteten Hochschulen zusammen nur noch in 0.1% der Fälle das Prädikat ‚ausreichend' als Gesamtabschlussnote vergeben, zwischen 1960 und 1979 sind es immerhin noch 0.8%. Auch die Anteile der ‚befriedigenden' (8.1% auf 4.1%) und selbst der ‚guten' (48.4% auf 39.0%) Abschlüsse sinken merklich. Lediglich die ‚sehr guten' Abschlüsse steigern ihren Anteil von 42.6% auf 56.8% und stellen damit mehr als die Hälfte aller Abschlüsse seit 1980 dar.

Abbildung 101: Verteilungen der Examensnoten in Psychologie 1960-1979 (links) und 1980-1997(rechts)

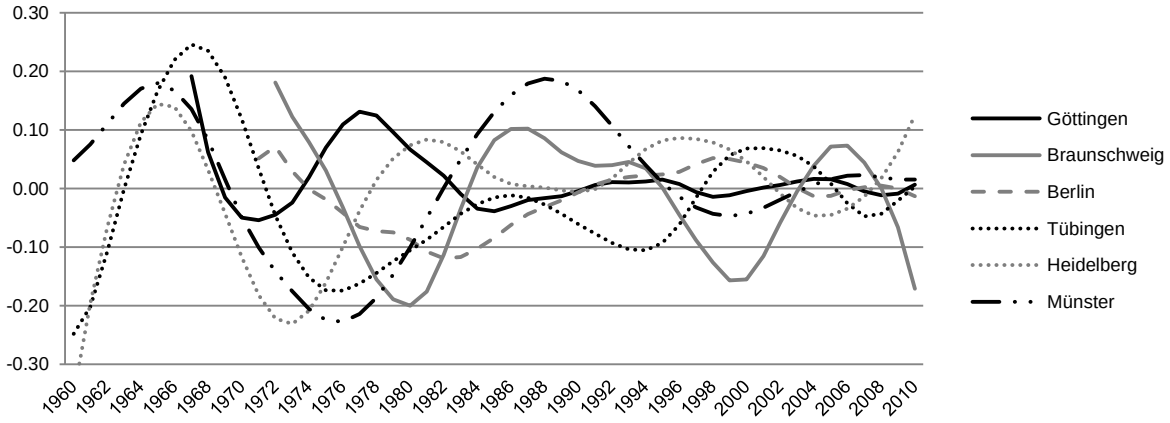


Es lässt sich in Tübingen (1973-1986) eine Plateauphase passend zur zeitlich parallelen Aufwärtsbewegung der zyklischen Komponente finden, wie dies auch in den anderen Studiengängen der Fall ist. Im Gegensatz zum üblichen Muster lassen sich jedoch in Göttingen (1991-1998), Münster (1998-2010) und Berlin (2003-2010) solche Phasen mit geringen Schwankungen (max. im Bereich von R=0.20 Noten Spannweite) finden, die nicht das Ergebnis einer zyklischen Aufwärtsbewegung, welche auf einen schwachen Abwärtstrend trifft, sind. In diesen Phasen scheint an den drei Hochschulen tatsächlich kaum Dynamik in den Noten zu liegen.

Die auf Studiengangebene vorzufindende starke Abwärtsbewegung der Noten zu Beginn der Zeitreihe lässt sich auf Hochschulebene an allen Universitäten außer in Göttingen wiederfinden. Die Stabilität der Noten aller Prüflinge im Diplomstudiengang Psychologie im Anschluss an diese starke Verbesserung ergibt sich aus dem Zusammenspiel der relativ stabilen Verläufe in Braunschweig, Tübingen und Münster und der sich ausgleichenden Bewegungen in Göttingen sowie Heidelberg in Richtung besserer und Berlin in Richtung schlechterer Noten.

Wie bereits in den Boxplots sichtbar, ist vor allem der Notenverlauf in Heidelberg durch eine starke Dynamik geprägt. Die Noten verlaufen in Zeiten allgemeiner Verbesserung parallel zum Studiengangmittel - nachdem die Verbesserung an allen anderen Hochschulen außer Göttingen abgeklungen ist, verbessern sich die Heidelberger Noten jedoch weiterhin, wodurch sie sich, anfangs noch über dem Durchschnitt aller Prüflinge liegend, dem Studiengangmittel zunächst annähern und es schließlich Ende der 1990er Jahre unterschreiten. Berlin und Münster nehmen den entgegengesetzten Verlauf einer anfänglichen Verbesserung (Berlin) bzw. Verschlechterung (Münster) des Notenniveaus im Vergleich zum Studiengangdurchschnitt, bevor sie sich durch einen jeweiligen Wendepunkt wieder aufeinander zu bewegen, um zum Ende der Reihe nahe dem mittleren Notenniveau zu konvergieren. Auch die Braunschweiger Noten liegen zum letzten Messzeitpunkt nah am Studiengangmittel, um das sie bereits seit den 1970ern relativ stabil verlaufen, wobei sie gemeinsam mit den Berliner Noten insgesamt die geringste Dynamik aufweisen.

Abbildung 102: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.3-LOWESS 0.9)



Volkswirtschaftslehre Diplom[76]

Nachdem in Biologie das bisher homogenste Notenniveau festgestellt wurde, liegt in VWL die größte Bandbreite an Durchschnittsnoten vor. Dies wird schon daran deutlich, dass die Skala der Abbildung der Differenzen zum Studiengangniveau im Vergleich zu den bisherigen Abbildungen um 0.2 Punkte weiter gefasst werden muss, um die Differenzen vollständig abzubilden.

Das liegt vor allem an der Notenentwicklung in Karlsruhe, die Mitte der 1990er einen starken Abfall Richtung Bestnoten vollzieht, bevor sie sich Anfang der 2000er wieder zurück in Richtung der Niveaus der anderen Hochschulen bewegt. Bis Ende der 1960er Jahre lässt sich zwischen unterem und oberem Ende ein Abstand ca. 0.5 Noten ablesen. In den 1970ern steigt dieser auf bis zu ca. 0.9 Noten an, sinkt in den 1980ern dann auf ca. 0.7 ab, steigt bis zum Ende der 1990er bis zu einem Wert von ca. 1.2 und endet schließlich bei ca. 0.6. Göttingen und Münster stellen dabei über den größten Zeitraum das obere Ende der Bandbreite dar, von ca. 1970-1990 begleitet von Tübingen. Karlsruhe, bis zu Beginn der 1990er begleitet von Berlin, begrenzt das Notenspektrum am unteren Ende, in Heidelberg treffen die Noten relativ konstant den Studiengangdurchschnitt und liegen damit zwischen den Extrempolen.

[76] Anmerkung zur Datenbasis: Für Karlsruhe sind von 2008-2010 keine Informationen in der Prüfungsstatistik enthalten, die Notendurchschnitte liegen für Karlsruhe nur von 1960 bis 2007 vor.

Abbildung 103: Durchschnittliche Abschlussnoten an den Hochschulen in VWL Diplom - Zeitverlauf (LOWESS 0.3)

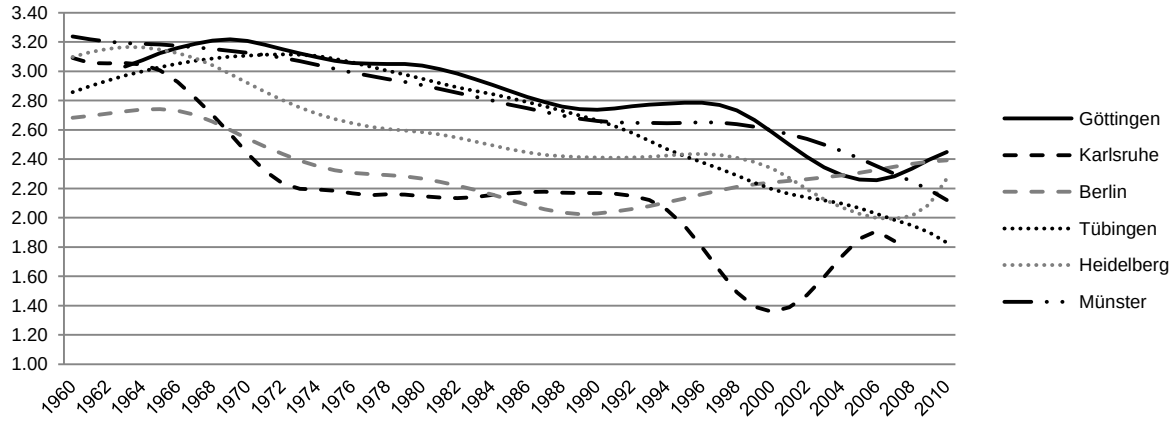
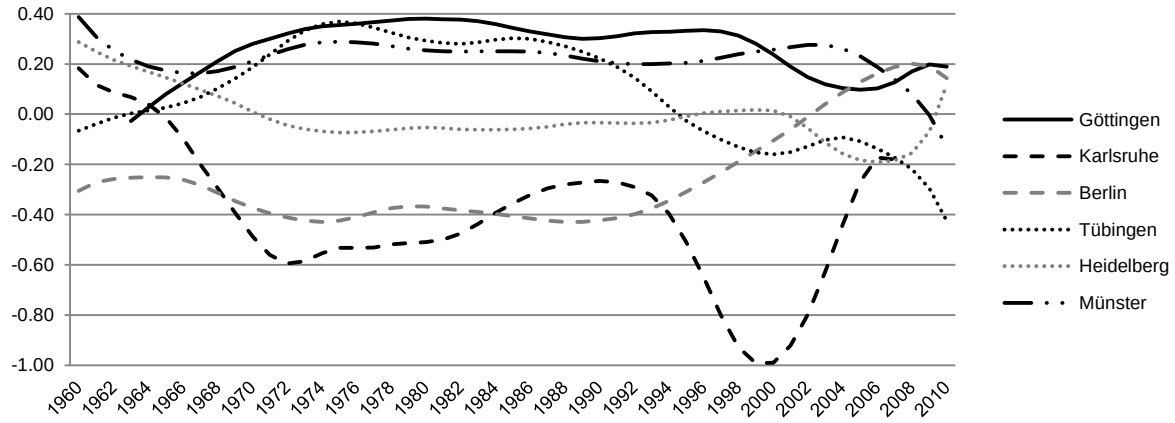


Abbildung 104: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Im Durchschnitt liegen die Karlsruher Noten 0.41 Noten unter dem Mittel aller Prüflinge, während die Göttinger Noten über den gesamten Zeitraum betrachtet 0.26 Noten schlechter als der Studiengangdurchschnitt sind. Im Zeitraum von 1963 bis 2007, in dem für alle sechs Hochschulen vollständige Daten vorliegen, sind die Karlsruher Noten im Durchschnitt 0.45 Noten besser als die aller VWL-Absolvent*innen, die Noten in Göttingen 0.27 Noten schlechter. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1960 gemittelt bei R=0.88 bzw. bei R=0.91 zwischen 1963-2007 und damit, wie in der Grafik bereits zu erkennen war, deutlich über den Werten aller bisher betrachteten Studiengänge.

Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen in beiden Zeiträumen 2003 (R=0.37), am größten 2001 (R=1.79). Die enorme Dynamik der Noten in Karlsruhe und Berlin sorgt für die große Differenz zwischen diesen Extremwerten, die die Aussagekraft des Durchschnitts der Spannweite über die Jahre im Vergleich zu den anderen Studiengängen klar einschränkt. Wird der Verlauf der Spannweite ohne die extrem niedrigen Noten in Karlsruhe in den 2000ern betrachtet, zeigt sich eine anfängliche Zunahme, gefolgt von einer kontinuierlichen Abnahme bis zum Ende der Zeitreihen.

Über jeweils 5-Jahresabschnitte gemittelt zeigt sich von ca. 1976-1995 eine grobe Notenhierarchie: Die besten Noten gibt es entweder in Karlsruhe oder Berlin. Heidelberg liegt in der Rangfolge in diesem Zeitraum stets auf dem dritten Platz, gefolgt von Münster und Tübingen auf den Rängen 4 und 5 (1991-1995 andersherum). In Göttingen sind die Noten immer am schlechtesten. Vor 1976 und nach 1996 findet sich Karlsruhe in der Regel auch auf dem ersten Platz, Göttingen kommt nur im Zeitraum bis 1966 über den fünften Rang hinaus.

Tabelle 38: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	BER	2.73	KA	2.70	KA	2.19	KA	2.15	KA	2.15	BER	2.02	BER	2.05	KA	1.40	KA	1.66	KA	1.82
2	TÜ	2.99	BER	2.73	BER	2.29	BER	2.34	BER	2.18	KA	2.20	KA	2.16	TÜ	2.22	TÜ	2.15	HD	1.91
3	GÖ	3.05	TÜ	3.05	HD	2.65	HD	2.68	HD	2.50	HD	2.39	HD	2.41	BER	2.25	HD	2.18	TÜ	1.96
4	KA	3.06	HD	3.11	MS	3.05	MS	2.94	MS	2.85	MS	2.66	TÜ	2.54	HD	2.48	BER	2.24	MS	2.25
5	HD	3.17	MS	3.20	GÖ	3.12	TÜ	3.01	TÜ	2.88	TÜ	2.74	MS	2.66	MS	2.63	GÖ	2.29	GÖ	2.33
6	MS	3.20	GÖ	3.24	TÜ	3.15	GÖ	3.06	GÖ	2.94	GÖ	2.76	GÖ	2.80	GÖ	2.74	MS	2.53	BER	2.37

Die einfaktorielle ANOVA offenbart entsprechend der im Zeitverlauf zwar schwankenden, aber durchgängig großen Spannweite zwischen bester und schlechtester Hochschule in 48 von 51 Jahren signifikante Unterschiede in den Durchschnittsnoten zwischen mindestens zwei Hochschulen, die kombinierte Überprüfung mit Kruskal-Wallis und Games-Howell Test ergibt 49 signifikante Unterschiede in 51 Jahren. Es unterscheiden sich die Noten in Münster und Berlin am häufigsten signifikant voneinander - in 38 von 51 Jahren, bei einer durchschnittlichen Fallzahl von 84 bzw. 72 Prüflingen pro Jahr im Vergleichszeitraum. Zwischen Berlin und Göttingen (33 von 48 Jahren) sowie Karlsruhe und Göttingen (29 von 44) besteht ein ähnliches Verhältnis.

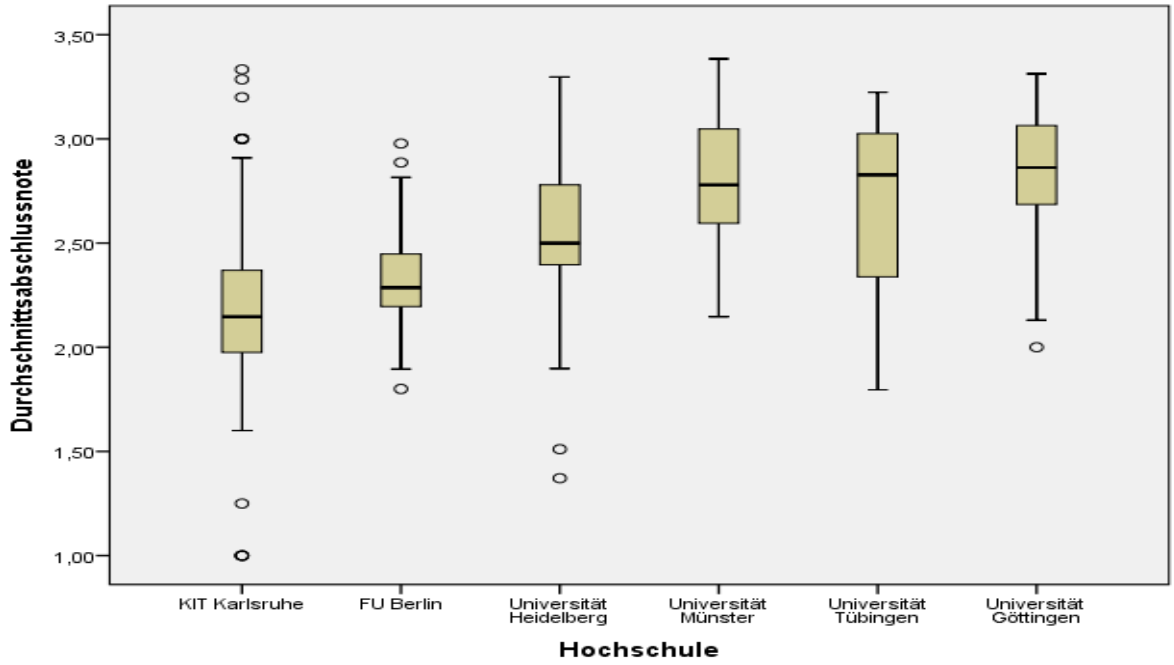
Die 15 Paarvergleiche beinhalten 20 Phasen durchgängig signifikanter Unterschiede, von denen sich vier als längerfristig (Göttingen/Karlsruhe, Göttingen/Berlin, Münster/Karlsruhe, Münster/Berlin) und zwei als mittelfristig (Tübingen/Karlsruhe, Tübingen/Berlin) stabil klassifizieren lassen. Die übrigen 14 Phasen weisen keine zeitliche Stabilität auf (siehe Anhang: Abb.A26).

Die Distanzen (Betragsfunktion der Differenz) zwischen den einzelnen Hochschulen reichen von 0.14 zwischen Göttingen und Münster bis 0.67 zwischen Karlsruhe und Münster. 10 der 15 Paarvergleiche liegen im Bereich von 0.21 bis 0.50 Noten Distanz, was deutlich über den Abständen im Notenniveau zwischen Hochschulen in den naturwissenschaftlichen Studiengängen und in Psychologie liegt. Die größte prozentuale Differenz zwischen den Notenniveaus lässt sich über die Jahre gemittelt zwischen Karlsruhe und Göttingen feststellen: Das Notenniveau in Karlsruhe erreicht durchschnittlich 74.9% des Göttinger Niveaus, sowohl über den gesamten Zeitraum seit 1960 als auch im Zeitraum von 1963-2007.

Die Verteilung der vergebenen Examensnoten über die Zeit zeigt, dass in Berlin das geringste Spektrum an Noten abgerufen wurde, in Heidelberg, Tübingen und Karlsruhe umfasst die Spannweite unter Nichtberücksichtigung der Ausreißer etwa eine halbe Note mehr. Werden die Ausreißer mit ein-

bezogen, wird vor allem die enorme Bandbreite vergebener Noten in Karlsruhe deutlich - dort reicht das Spektrum an Durchschnittsnoten zwischen 1960 und 2010 von $\bar{x}$=1.00 bis $\bar{x}$=3.33 und damit über 2 Noten weit. Die im Vergleich zu den anderen Studiengängen hohe Spannweite an allen Hochschulen deckt entsprechend der schlechteren Noten dort v.a. in Münster und Göttingen nur den Bereich oberhalb der Marke von $\bar{x}$=2.00 ab, die mittleren 50% der Noten dort liegen in einem Bereich des Notenspektrums, an das die mittleren 50% der Noten in Karlsruhe und Berlin nicht heranreichen. In Berlin bleibt $\bar{x}$=3.00 durch das insgesamt gute Notenlevel die obere Begrenzung, was an keiner anderen Hochschule der Fall ist.

Abbildung 105: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1960-2010



Die Noten streuen über die Zeit an allen Standorten außer Karlsruhe ähnlich stark. In Karlsruhe ist die Streuung der Streuung sehr hoch, so dass der vergleichsweise niedrige Durchschnittswert der Standardabweichung über die Jahre gemittelt kaum aussagekräftig ist. Die Stärke der durchschnittlichen Streuung unterscheidet sich ebenfalls mit Ausnahme von Karlsruhe nur geringfügig. Die Werte liegen ungefähr im gleichen Bereich wie in Chemie und damit niedriger als in den beiden Mathematik-Studiengängen, aber höher als in Biologie und Psychologie. Da der Zeitraum, in dem für alle Hochschulen Daten vorliegen nur sechs Jahre vom Gesamtzeitraum abweicht, ergeben sich bei der Betrachtung keine nennenswerten Differenzen.

Tabelle 39: Streuung der Noten an den Hochschulen im Diplomstudiengang VWL

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1960-2010	1963-2007	1960-2010	1963-2007
Karlsruhe (bis 2007)	0.52	0.50	0.25	0.24
Münster	0.61	0.61	0.08	0.07
Tübingen	0.63	0.62	0.10	0.09
Heidelberg	0.64	0.62	0.09	0.07
Göttingen (ab 1963)	0.65	0.66	0.08	0.08
Berlin	0.67	0.68	0.08	0.08

Die Entwicklung der Noten an den einzelnen Hochschulen im Verlauf seit 1960 geht vor allem in eine Richtung - in Richtung besserer Noten. In Karlsruhe ist die Abwärtsbewegung trotz zwischenzeitlichem Aufschwung in den 2000ern Jahren insgesamt am stärksten: Das mittlere Notenniveau der letzten 10 Jahre in der Reihe liegt um knapp 1.5 Noten unter dem der ersten 10, was eine sehr starke Verbesserung bedeutet. Die durchschnittliche jährliche Veränderung beträgt seit 1960 -0.033 Noten pro Jahr, was eine Differenz von -1.58 zwischen letztem (2007) und erstem (1960) Wert zufolge hat. Im stärksten Trendbereich 1964-1998 ergibt sich eine Verbesserung von -2.33 Noten bei einer durchschnittlichen Veränderung von -0.069 Noten pro Jahr. Auffällig ist, dass sich die Verbesserung in zwei relativ kurzen Zeiträumen von Mitte der 1960er bis Anfang der 1970er und im Laufe der 1990er Jahre vollzieht, während in der Zwischenzeit ein relativ stabiles Niveau herrscht.

Auch in Tübingen, Münster, Heidelberg und Göttingen verbessern sich die Noten langfristig deutlich - allerdings fast kontinuierlich, mit den inzwischen bereits bekannten zyklischen Schwankungen von 10-20 Jahren Dauer (Abb.103 und 106). In Heidelberg weist das mittlere Notenniveau der 2000er immerhin noch eine Verbesserung von -1.12 Noten im Vergleich zu den 1960ern auf. In Tübingen, Göttingen und Münster nimmt die langfristige Verbesserung in dieser Reihenfolge ab, liegt aber auch dort immer noch bei mindestens 80% einer ganzen Note. In Berlin hingegen stoppt die Verbesserung zu Beginn der 1990er und es beginnt eine Phase der Verschlechterung, die bis zum letzten Messzeitpunkt anhält, aber bis dahin nicht wieder das Ausgangsniveau der 1960er Jahre erreicht. Langfristig ergibt sich dort entsprechend das geringste Ausmaß an Verbesserung und das früheste Ende des stärksten Trendbereichs, der an allen Hochschulen in den 1960ern beginnt, an den anderen fünf Hochschulen aber mindestens bis zum Ende der 1990er anhält, an drei von ihnen sogar bis zum Ende der 2000er, wodurch die - immer noch hohen - jährlichen Veränderungen im Durchschnitt unterhalb der Werte in Psychologie liegen.

Tabelle 40: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang VWL 1960-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Karlsruhe (bis 2007)	-1.49	-1.58	-0.033	-2.33 (1964-1998)	-0.069
Heidelberg	-1.12	-0.82	-0.016	-1.93 (1964-2007)	-0.045
Tübingen	-0.96	-1.06	-0.021	-1.42 (1971-2010)	-0.037
Göttingen (ab 1963)	-0.87	-0.67	-0.014	-1.31 (1967-2003)	-0.036
Münster	-0.81	-1.23	-0.025	-1.23 (1960-2009)	-0.025
Berlin	-0.44	-0.26	-0.005	-1.18 (1963-1990)	-0.044

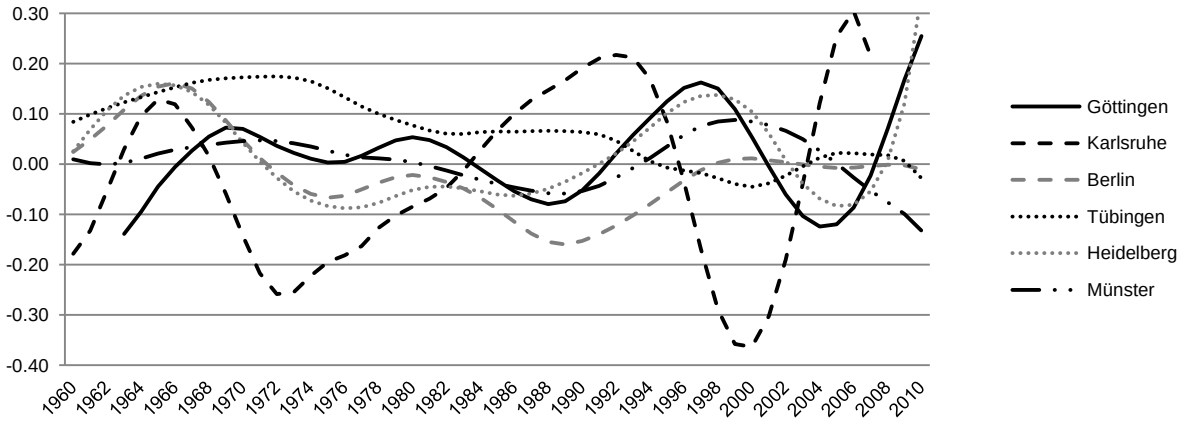
Tabelle 41: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang VWL 1963-2007

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Karlsruhe	-1.28	-1.17	-0.026	-2.33 (1964-1998)	-0.069
Tübingen	-0.97	-0.80	-0.018	-1.39 (1971-1998)	-0.052
Heidelberg	-0.96	-1.76	-0.040	-1.93 (1964-2007)	-0.045
Göttingen	-0.77	-0.84	-0.019	-1.31 (1967-2003)	-0.036
Münster	-0.63	-0.94	-0.021	-0.94 (1963-2007)	-0.021
Berlin	-0.44	-0.57	-0.013	-1.18 (1963-1990)	-0.044

Es lassen sich mit Ausnahme von Karlsruhe an allen Universitäten passend zu zeitlich parallelen Aufwärtsbewegungen der zyklischen Komponenten Plateauphasen finden: In Tübingen (1964-1977), Göttingen (1974-1983) und Heidelberg (1991-1998) ist jeweils eine, in Münster (1964-1972 und 1990-2000) und Berlin (1974-1983 und 1997-2006) sind jeweils zwei Perioden (vermeintlicher) Konstanz in den Notenbewegungen zu finden. Insgesamt bietet sich auf Hochschulebene damit ein Bild, das dem auf Studiengangebene entspricht: Bis auf die zu Beginn der 1990er beginnende Aufwärtsphase in Berlin findet sich überall die langfristige, fast durchgängige, wenn auch in zyklischen Bewegungen verlaufende starke Verbesserung der Noten, die sich auch im Mittel aller VWL-Noten zeigt. Entsprechend den parallelen Verläufen ist auch kaum Dynamik im Verhältnis der Notenniveaus untereinander zu verzeichnen. Lediglich die starken Schwankungen in Karlsruhe und die Verbesserung in Berlin wirken als Veränderungen der Hochschulrelationen.

Im Vergleich zum Studiengangmittel, dem das Heidelberger Niveau in etwa entspricht, zeigt sich außerdem eine nach anfänglicher Absetzung in Richtung schlechterer Noten erfolgende Annäherung der Tübinger Noten im Zeitverlauf, die schließlich sogar unter dem Durchschnitt endet, was mit der stärkeren Verbesserung der Noten dort im letzten erfassten Zeitabschnitt zu erklären ist.

Abbildung 106: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.3-LOWESS 0.9)



Betriebswirtschaftslehre Diplom

In BWL stehen grundsätzlich Daten aus fünf Universitäten zur Verfügung. Da die vorhandenen Daten für Karlsruhe aber nur für eine Zeitreihenlänge von 18 Jahren (1964-1981) ausreichen, werden sie im Folgenden nicht berücksichtigt. Von ca. 1980 bis in die Mitte der 2000er Jahre bilden Göttingen und Münster das obere Ende der Notenskala, vor 1980 übernimmt Münster diese Rolle alleine. Die zunächst relativ große Bandbreite von ca. 0.6 Noten Unterschied zwischen Münster (und Göttingen) am oberen und Berlin am unteren Ende nimmt ab den 1970ern Jahren immer weiter ab, bis die besten und schlechtesten Noten 2010 nur noch ca. 0.15 Noten auseinanderliegen (Abb.107 und 108). Durch die vergleichsweise hohe Anzahl der Berliner Absolvent*innen (durchschnittlich n=180 pro Jahr seit 1960) im Vergleich zu Göttingen und Münster (zusammen n=255 pro Jahr) kombiniert mit den ver-

gleichsweise guten Noten in Berlin liegen Göttingen (und Münster ab ca. 1980) deutlich geringer über dem Studiengangmittel als Berlin darunter liegt. Tübingen bewegt sich über den kurzen erfassten Zeitraum zwischen diesen Extrempolen.

Abbildung 107: Durchschnittliche Abschlussnoten an den Hochschulen in BWL Diplom - Zeitverlauf (LOWESS 0.3)

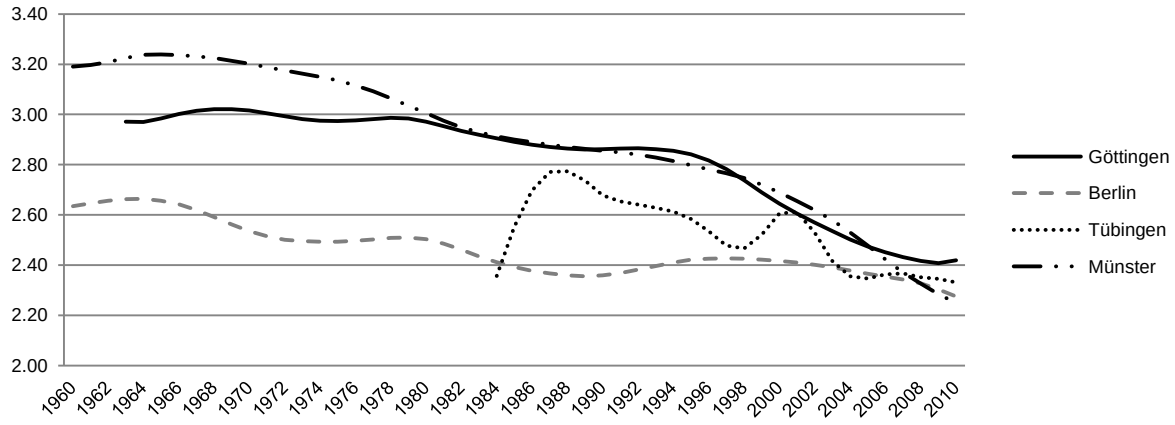
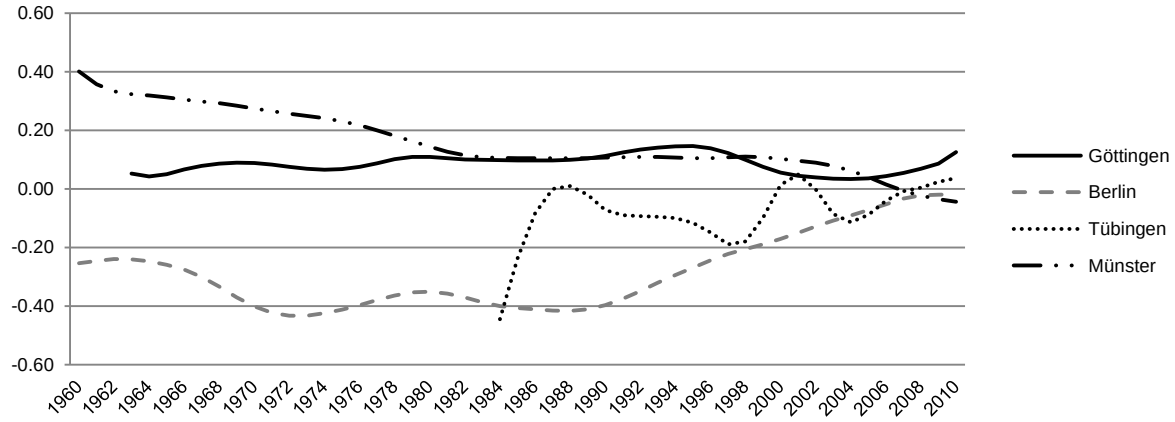


Abbildung 108: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Im Mittel über den gesamten Zeitraum sind die Noten der Münsteraner Prüflinge durchschnittlich 0.16 Noten schlechter als die aller Prüflinge, in Berlin sind sie durchschnittlich 0.28 Noten besser. Wird nur der Zeitraum betrachtet, in dem für alle Hochschulen im sample Werte vorliegen (1984-2010) schneiden die Göttinger Absolvent*innen mit 0.09 Noten über dem Durchschnitt am schlechtesten, die Berliner auch hier (0.23 Noten unter dem Studiengangmittel) am besten ab. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1960 gemittelt bei R=0.49 bzw. bei R=0.39 zwischen 1984-2010 und damit im Mittel so wie es die in der Grafik sichtbare zunächst große, dann sinkende Spanne erwarten lässt. Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen über den gesamten Zeitraum gesehen 2007 (R=0.04) für beide Zeiträume. Am größten ist diese Differenz 1973 (R=0.92) bzw. 1988 (R=0.73). Vor 1973 steigt die maximale Differenz im Notenniveau zwischen den Hochschulen zunächst einmal leicht an.

Die Berechnung der 5-Jahres-Mittel zeigt eine klare Notenhierarchie: Von 1961 bis 1980 werden die besten Noten in Berlin, die schlechtesten in Münster vergeben, Göttingen liegt dazwischen. Von 1981 bis 2005 reiht sich Tübingen auf Platz 2 ein, die Göttinger Noten sind in diesem Zeitraum in der Regel die schlechtesten.

Tabelle 42: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	BER	2.68	BER	2.62	BER	2.41	BER	2.56	BER	2.44	BER	2.34	BER	2.39	BER	2.45	BER	2.39	BER	2.32
2	GÖ	2.98	GÖ	3.04	GÖ	2.95	GÖ	3.00	GÖ	2.91	TÜ	2.76	TÜ	2.62	TÜ	2.51	TÜ	2.47	MS	2.32
3	MS	3.22	MS	3.26	MS	3.15	MS	3.08	MS	2.92	MS	2.86	MS	2.85	MS	2.72	GÖ	2.54	TÜ	2.35
4											GÖ	2.87	GÖ	2.88	GÖ	2.73	MS	2.60	GÖ	2.40

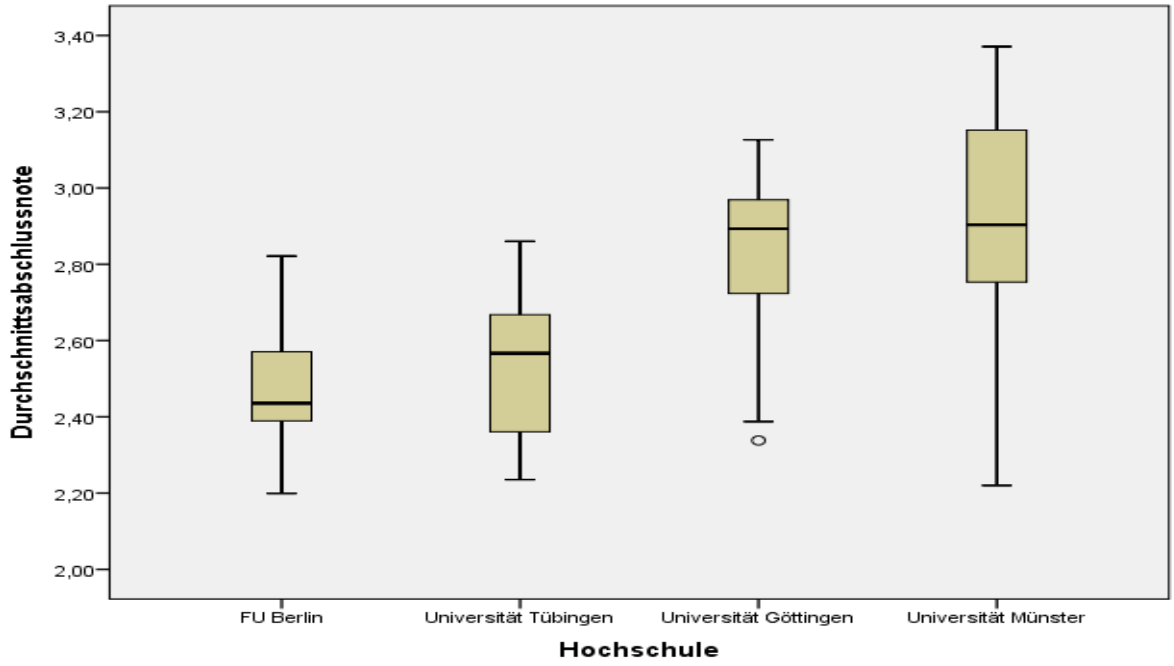
Die ANOVA zeigt, dass die Unterschiede zwischen den Hochschulen trotz Annäherung der Niveaus am Ende der Reihen in 48 von 51 Jahren signifikant sind. Die Überprüfung der Ergebnisse anhand des Kruskal-Wallis und des Games-Howell Tests ergibt mit 47 signifikanten Werten ein Jahr weniger als signifikant aus. Am häufigsten unterscheiden sich die Noten in Berlin und Münster signifikant voneinander - in 43 von 51 Jahren bei einer durchschnittlichen Zahl von n=180 bzw. 288 Prüflingen pro Jahr im Vergleichszeitraum. Ein ähnliches Verhältnis besteht zwischen Berlin und Göttingen (35 von 48 Jahren bei n=186 bzw. n=219).

Die sechs Paarvergleiche umfassen sieben Perioden durchgängig signifikanter Differenzen. Zwischen Berlin und Münster sind die Differenzen im Notenniveau als langfristig stabil einzuordnen, zwischen Berlin und Göttingen immer noch als längerfristig stabil. Tübingen weist zu allen drei anderen Hochschulen Phasen kurzfristig stabiler Unterschiede auf (siehe Anhang: Abb.A27).

Über den gesamten Zeitraum gemittelt liegt die niedrigste Distanz (also die Betragsfunktion der Differenz zwischen den Notenniveaus) bei 0.18 (zwischen Tübingen und Münster), die höchste bei 0.45 (zwischen Berlin und Münster) Noten. Für die restlichen 4 Paarvergleiche liegen die Werte bei 0.11, 0.20, 0.21 und 0.37. Die größte prozentuale Differenz zwischen den Notenniveaus lässt sich über die Jahre gemittelt zwischen Berlin und Münster feststellen. Das Notenniveau in Berlin erreicht durchschnittlich 85.4% des Münsteraner Niveaus. Für den Zeitraum mit allen Hochschulen im sample liegt die höchste Differenz zwischen Berlin und Göttingen - in Berlin werden 88.6% des Göttinger Niveaus erreicht.

Die Verteilung über die Zeit offenbart eine klare Zweiteilung: Die mittleren 50% der Noten in Göttingen und Münster decken einen komplett anderen Bereich des Notenspektrums ab, als die mittleren 50% in Berlin und Tübingen. Die Spannweite der Noten reicht in Göttingen allerdings bis 0.20 Noten über die untere Begrenzung der Verteilungen an den beiden Hochschulen mit den besseren Noten, in Münster reicht sie sogar gleich weit, was bedeutet, dass die Bandbreite der dort erteilten Abschlussnoten fast doppelt so groß ist.

Abbildung 109: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1960-2010



Hinsichtlich der über die Jahre gemittelten Streuung lassen sich, abgesehen von einem vergleichsweise niedrigen Wert für Münster, keine großen Unterschiede finden. Die Werte liegen für beide betrachteten Zeiträume nah beieinander und insgesamt im mittleren Bereich der untersuchten Studiengänge. Die Streuung ist über die Jahre hinweg ebenfalls einheitlich und auch sehr stabil, die Wertelie gen allesamt im niedrigen Bereich.

Tabelle 43: Streuung der Noten an den Hochschulen im Diplomstudiengang BWL

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1960-2010	1984-2010	1960-2010	1984-2010
Münster	0.57	0.56	0.05	0.04
Tübingen (ab 1984)	0.61	0.61	0.07	0.07
Göttingen (ab 1963)	0.63	0.61	0.06	0.04
Berlin	0.65	0.61	0.07	0.07

Die Entwicklung der Noten an den einzelnen Hochschulen seit 1960 im Verlauf zeigt eine in Göttingen in Zyklen verlaufende Abwärtsbewegung, die mit dem Anfang der 1980er an Stärke zunimmt und zu einer Verbesserung von gut einer halben Note in den letzten 10 Jahren gegenüber den ersten 10 Jahren der Zeitreihe führt. In Berlin setzt die erste Verbesserung einige Jahre früher und auf niedrigerem Niveau ein als in Göttingen und Münster. Der ebenfalls von Zyklen begleitete Abwärtsverlauf nimmt ab Mitte der 1980er Jahre in seiner Stärke ab, wodurch insgesamt auch eine geringere Verbesserung - ca. eine Drittelnote im Vergleich der 10-Jahresmittel - als in Göttingen und Münster entsteht. Münster weist nicht nur die stärkste Entwicklung zu besseren Noten auf, sondern auch die konstanteste - aber auch die bereits vor der Trendbereinigung ablesbaren üblichen 10-20 jährigen Zyklen (Abb.107 und 110). In Tübingen ist für den relativ kurzen Zeitraum, in dem Daten vorliegen, ebenfalls eine Verbesserung zu beobachten, auch eine zyklische Komponente ist zumindest ansatzweise erkennbar. Dass zunächst ein steiler Anstieg zu Beginn der Reihe zu sehen ist, ist darauf zu-

rückzuführen, dass der erste Wert 1984 äußerst niedrig liegt - da auch die beiden nächsten Werte noch niedriger sind als die folgenden Höchstwerte Mitte der 1980er entsteht bei der Glättung der Daten der Eindruck, es würde zunächst eine Phase deutlichen Anstiegs vorliegen - tatsächlich entsteht dieser Eindruck aber nur durch die hohe Differenz zwischen dem erstem und den folgenden Werten. Als Konsequenz zeigt sich in der absoluten Veränderung sogar ein leichter Anstieg.

Die Differenz der 10-Jahresmittel, die in diesem Fall aufgrund der geringen Länge der Zeitreihe beinahe einer Zweiteilung der Daten gleichkommt, korrigiert diesen Eindruck aber und zeigt eine etwas geringere Verbesserung als in Berlin. Im stärksten Trendbereich, der in Göttingen etwas später einsetzt als in Münster und Berlin, dort aber früher endet, erreicht Tübingen im wesentlich kürzeren Zeitraum sogar die höchste durchschnittliche Veränderung, vor Münster und Göttingen. Berlin weist auch hier die geringsten Verbesserungswerte auf. Am Vergleichszeitraum, in dem für alle Hochschulen Werte vorliegen, ist ersichtlich, dass der Abwärtstrend in Berlin zunehmend abschwächt: Das mittlere Niveau der 2000er Jahre liegt nur noch minimal (-0.01) Noten unter dem mittleren Niveau von 1984-1993, während in Göttingen und Münster auch hier noch Verbesserungen von ca. 40% einer ganzen Note zu verzeichnen sind.

Tabelle 44: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang BWL 1960-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Münster	-0.79	-1.05	-0.021	-1.15 (1960-2008)	-0.024
Göttingen (ab 1963)	-0.54	-0.57	-0.012	-0.79 (1969-2008)	-0.020
Berlin	-0.31	-0.40	-0.008	-0.62 (1965-2002)	-0.017
Tübingen (ab 1984)	-0.26	+0.05	+0.002	-0.62 (1988-2009)	-0.030

Tabelle 45: Kennzahlen - Notenentwicklung an den Hochschulen im Diplomstudiengang BWL 1984-2010

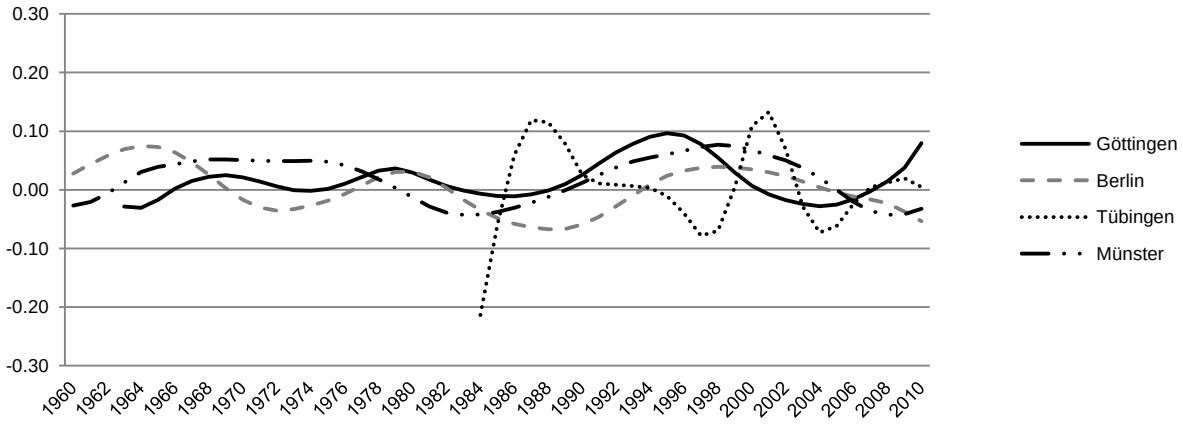
Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Münster	-0.41	-0.63	-0.024	-0.73 (1984-2008)	-0.030
Göttingen	-0.39	-0.44	-0.017	-0.60 (1994-2008)	-0.043
Tübingen	-0.26	+0.05	+0.002	-0.62 (1988-2009)	-0.030
Berlin	-0.01	-0.28	-0.011	-0.36 (1984-2002)	-0.020

Die Betrachtung der Notenentwicklung an den einzelnen Hochschulen entspricht dem Verlauf der Noten auf Studiengangebene. An allen vier betrachteten Universitäten verläuft das Absinken des Notenniveaus in den Wellen, die sich auch über alle Prüflinge gemittelt zeigen. Die zunehmende Trendstärke in Göttingen gegen Anfang der 1980er Jahre und die Einbindung der Tübinger Noten in den Gesamtdurchschnitt ab 1984 entspricht dem zunehmenden Abwärtstrend in diesem Bereich der Zeitreihe auf Studiengangebene.

Ausgeprägte Plateauphasen lassen sich in Göttingen (1983-1994), Münster (1989-1995) und Berlin (1989-2001) finden, wiederum parallel zu aufwärtsstrebenden Zyklen. Die Noten in Münster weisen in ihrem Verlauf die größte absolute Veränderung auf. Durch die starke Verbesserung der Noten in Göttingen ab Mitte der 1980er und die parallele Abschwächung des Trends in Berlin nähert sich das

mittlere Niveau aller Prüflinge zunehmend dem Berliner Niveau an, so dass die Noten dort nach Jahrzehnten unter dem Durschnitt, nahezu das Studiengangniveau treffen.

Abbildung 110: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.3-LOWESS 0.9)



Soziologie Magister/Diplom[77]

Inklusive der Berliner Noten zeigt sich eine ab- und anschließend wieder zunehmende Spannweite der Durchschnittsnoten zwischen den Hochschulen. Die Berliner Noten stellen über den größten Zeitraum das untere Ende des Spektrums dar, Heidelberg und Göttingen, bis in die 1990er, und anschließend Tübingen bilden die obere Begrenzung. In Münster befinden sich die Noten in diesem Fall zwischen den drei anderen Magisterstandorten ober- und den Berliner Noten unterhalb des Studiengangmittels. Werden ausschließlich die Magisterstudiengänge betrachtet, nimmt Münster anstelle von Berlin den Platz als Hochschule mit den über den größten Zeitraum besten Noten ein. Mit Berlin entwickelt sich der Abstand zwischen am besten und am schlechtesten bewertender Hochschule von ca. 0.2 Noten über bis zu 0.9 Noten Differenz in den 1980ern zu schließlich etwas unter 0.4 Noten. Ohne Berlin ist der Abstand erheblich geringer, entwickelt sich von knapp unter 0.2 Noten Differenz

[77] Anmerkungen zur Datenbasis: In Soziologie stehen Daten aus fünf Universitäten zur Verfügung. An vier Standorten wurde im betrachteten Zeitraum im Fach nur der Magisterabschluss über genügend Jahre angeboten, um eine sinnvolle Analyse der Daten durchführen zu können. In Berlin hingegen wurde durchgängig der Diplomabschluss in Soziologie angeboten. Das Notenniveau in Berlin unterscheidet sich über einen weiten Zeitraum deutlich von dem Niveau der Hochschulen, an denen der Magisterabschluss erhoben wurde. Aufgrund der hier verwendeten Daten alleine kann nicht entschieden werden, ob es sich bei der Differenz zwischen den Berliner Noten und denen der übrigen Universitäten um eine hochschulspezifisch oder um eine abschlussspezifisch bedingte Differenz handelt.
Der in Abschnitt 8.1.1 durchgeführte Vergleich der Studiengangmittel der Stichprobe mit denen aller westdeutschen Hochschulen spricht zumindest für den erfassten Zeitraum von 1996 bis 2010 gegen einen abschluss- und für einen hochschulspezifischen Effekt, da sich die Durchschnittsnoten in Magister- und Diplomstudiengang kaum unterscheiden und das Mittel aller westdeutschen Prüfungsergebnisse im Diplom deutlich höher liegt als das Berliner Niveau. Was die (größere) Differenz zwischen den vier Magisterstandorten und Berlin vor 1996 betrifft, lassen sich jedoch keine Zusatzinformationen finden, die entsprechende Schlussfolgerungen ermöglichen, weshalb die deskriptive Analyse auf Hochschulebene zweimal, sowohl für alle Hochschulen im Vergleich als auch nur für den Magisterstudiengang, durchgeführt wird.
Da für den Magisterstudiengang vor 1969 keine ausreichenden Fallzahlen vorliegen (hochschulübergreifend maximal n=10 in 1967), werden der Vergleichbarkeit halber beide Analysewege erst ab diesem Jahr durchgeführt.

über einen Maximalbereich von ca. 0.5 Noten Unterschied hin zu ebenfalls etwas unter 0.4 Punkten Differenz.

Abbildung 111: Durchschnittliche Abschlussnoten an den Hochschulen in Soziologie - Zeitverlauf (LOWESS 0.3)

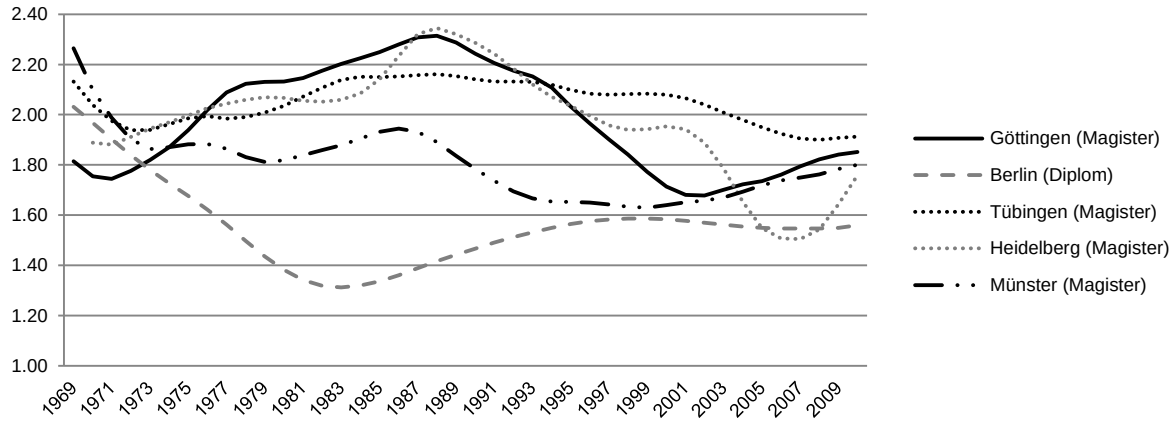


Abbildung 112: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt – Alle Hochschulen (LOWESS 0.3)

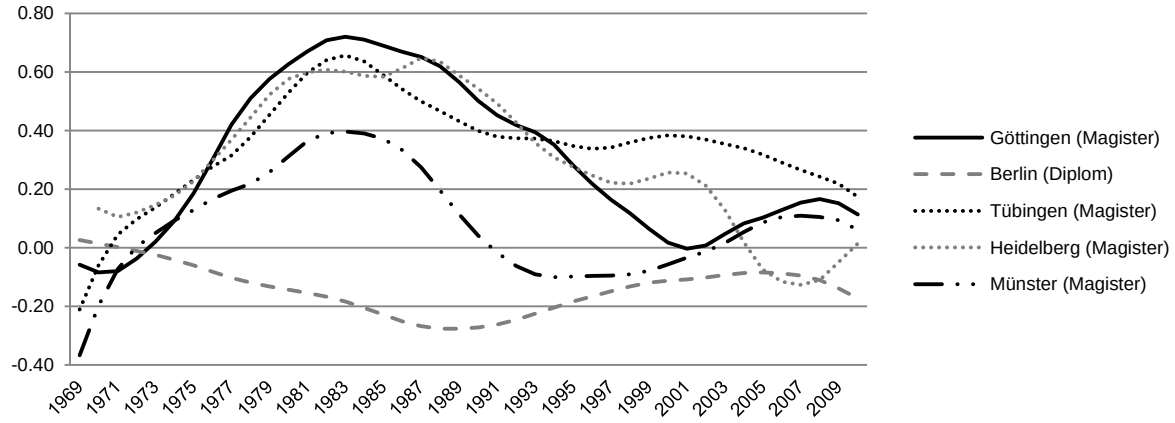
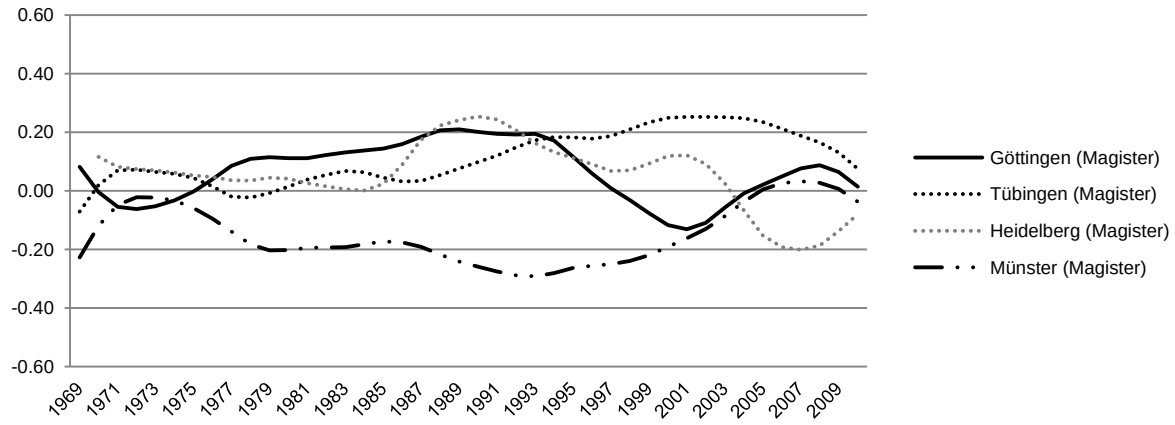


Abbildung 113: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt – Nur Magister (LOWESS 0.3)



Im Mittel über den gesamten Zeitraum seit 1969 sind die Noten der Berliner Prüflinge durchschnittlich 0.14 Noten besser als die aller Prüflinge, in Tübingen sind sie durchschnittlich 0.35 Noten schlechter. Wird nur der Zeitraum betrachtet, in dem für alle Hochschulen im sample Werte vorliegen (1970-2010) liegen die Tübinger Absolvent*innen mit 0.37 Noten über dem Durchschnitt, die Berliner auch hier mit 0.14 Noten darunter. Nur das Mittel der Magisterprüflinge betrachtet, liegen

(in beiden Zeiträumen) die Tübinger Absolvent*innen mit 0.12 Noten über diesem und die Münsteraner 0.14 darunter. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt in beiden Zeiträumen gemittelt bei R=0.75 mit und bei R=0.55 ohne Berlin. Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen über den gesamten Zeitraum gesehen mit Berlin 1970 (R=0.24), am größten 1988 (R=1.18) für beide Zeiträume.

Diese Werte entsprechen der Gesamtentwicklung, die Spannweite nimmt bis in die 1980er zu, bevor sie wieder bis zum Ende der Reihen abnimmt. Ohne Berlin ist diese Differenz 1974 (R=0.14) am geringsten und 1972 am größten (R=1.13). Nach dieser starken Schwankung zu Beginn der Reihen nehmen die kurzfristigen Ausschläge zwar in der Höhe etwas ab, bis zum Ende der Zeitreihen schwankt die maximale Differenz aber dennoch im Bereich von ca. einer halben Note, ohne dass eine Tendenz zur Zu- oder Abnahme besteht. Die Berechnung der 5-Jahres-Mittel zeigt keine klare Notenhierarchie. Lediglich, dass von 1976 bis 2005 die besten Noten in Berlin, von 1976 bis 2000 die zweitbesten (ohne die Berücksichtigung Berlins die besten) in Münster sowie von 1996 bis 2010 die schlechtesten Noten in Tübingen vergeben wurden, lässt sich hier erkennen.

Tabelle 46: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

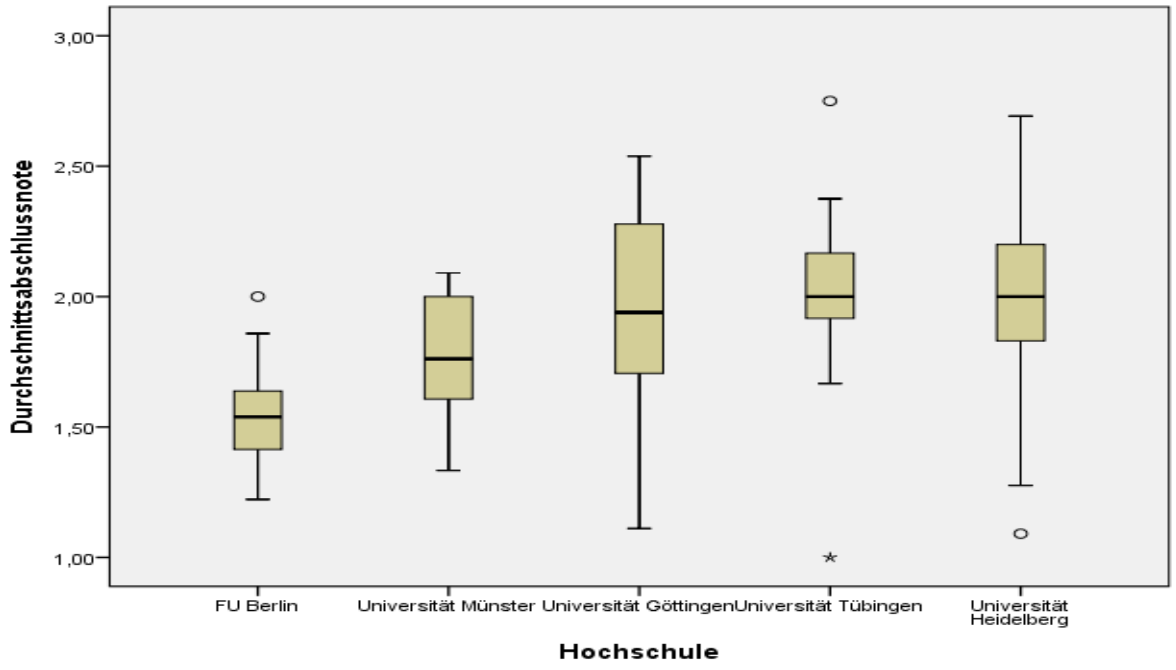
	1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	GÖ	1.75	BER	1.48	BER	1.26	BER	1.43	BER	1.53	BER	1.61	BER	1.57	HD	1.50
2	BER	1.79	MS	1.82	MS	1.85	MS	1.91	MS	1.68	MS	1.61	GÖ	1.59	BER	1.54
3	MS	1.90	TÜ	1.98	HD	2.05	TÜ	2.14	TÜ	2.12	GÖ	1.81	MS	1.67	MS	1.77
4	HD	1.94	HD	2.03	TÜ	2.13	GÖ	2.35	HD	2.12	HD	1.87	HD	1.83	GÖ	1.87
5	TÜ	1.95	GÖ	2.19	GÖ	2.13	HD	2.42	GÖ	2.19	TÜ	2.10	TÜ	2.02	TÜ	1.89

Die ANOVA zeigt, dass die Unterschiede zwischen den Hochschulen bei Berücksichtigung Berlins in 31 von 42 Jahren signifikant sind, ansonsten nur in 10 Jahren. Die Überprüfung der Ergebnisse anhand des Kruskal-Wallis und des Games-Howell Tests ergibt mit 27 signifikanten Werten für den ersten und neun für den zweiten Fall vier Jahre bzw. ein Jahr weniger als signifikant aus. Am häufigsten unterscheiden sich die Noten in Berlin und Heidelberg signifikant voneinander - in 15 von 41 Jahren bei einer durchschnittlichen Zahl von n=97 bzw. n=15 Prüflingen pro Jahr im Vergleichszeitraum. Wird Berlin außen vor gelassen, findet sich die höchste Anzahl signifikant differenter Jahre zwischen Heidelberg und Münster. Hier sind es nur fünf von 41, bei allerdings niedrigen durchschnittlichen Fallzahlen von n=15 bzw. n=21 pro Jahr. Perioden, in denen die signifikanten Differenzen einen längeren Zeitraum überbrücken gibt es außer einer dreijährigen Phase zwischen Heidelberg und Münster nur unter Beteiligung von Berlin. Die Noten dort weisen zu denen aller anderen Standorte durchgängig signifikante Unterschiede auf, die allesamt im Bereich zwischen vier und sechs Jahre liegen.

Über den gesamten Zeitraum gemittelt, liegt die niedrigste Distanz (die Betragsfunktion der Differenz zwischen den Notenniveaus) bei 0.27 (zwischen Tübingen und Heidelberg), die höchste bei 0.54 (zwi-

schen Berlin und Tübingen) Noten. Sechs der übrigen acht Paarvergleiche liegen bei Werten von 0.30 bis 0.32, was einen relativ einheitlichen Abstand bedeutet. Ohne Berlin ergeben sich sechs Paarvergleiche, deren Distanzen allesamt im Bereich von 0.27 bis 0.32 liegen. Die größte prozentuale Differenz zwischen den Notenniveaus besteht über die Jahre gemittelt zwischen Berlin und Tübingen. Das Notenniveau in Berlin erreicht durchschnittlich 77.9% (77.0% für 1970-2010) des Tübinger Niveaus. Ohne Berlin liegt die höchste Differenz zwischen Münster und Tübingen - in Münster werden 88.5% (88.8% für 1970-2010) des Tübinger Niveaus erreicht. Die Spannweiten der Noten in Bezug auf ihre Verteilung innerhalb des Zeitraums von 1969 bis 2010 sind in Göttingen und Heidelberg ohne Berücksichtigung der Ausreißer mit weit über einer ganzen Note fast doppelt so groß wie an den übrigen drei Standorten. Die mittleren 50% der Noten sind in Göttingen am weitesten verteilt und umfassen den Interquartilsabstand in Heidelberg und Tübingen sowie einen Großteil dieses Bereichs in Münster. Die Berliner Noten dagegen heben sich in ihrer Verteilung über die Zeit deutlich zum Besseren ab, hier gibt es nur eine geringe Überschneidung der Box mit den Münsteraner mittleren 50%, ansonsten mit keiner der anderen Boxen.

Abbildung 114: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1969-2010



Die über die Jahre gemittelten Standardabweichungen und deren Standardabweichungen lassen erkennen, dass die Noten über die Zeit unterschiedlich stark streuen. In Berlin ist die Streuung der mittleren Standardabweichung recht niedrig, in Tübingen und Heidelberg ist sie etwa dreifach höher als dort. Göttingen und Münster erreichen immer noch etwa doppelt so hohe Abweichungen der Streuung im Zeitverlauf wie Berlin. Auch die Stärke der durchschnittlichen Streuung unterscheidet sich deutlich, mit einem niedrigen Wert für Berlin und vergleichsweise hohen Werten an den anderen Standorten, innerhalb derer sich Tübingen noch einmal etwas nach unten und Göttingen etwas

nach oben absetzt. Die gemittelte Streuung lässt sich in etwa auf dem Niveau von VWL und BWL einordnen, sie liegt niedriger als in den beiden Mathematikstudiengängen, allerdings deutlich höher als in Biologie und Psychologie. Der zweite Vergleichszeitraum mit allen Hochschulen im sample unterscheidet sich nur um ein Jahr vom Gesamtzeitraum, weshalb nur minimale Unterschiede auftreten.

Tabelle 47: Streuung der Noten an den Hochschulen im Studiengang Soziologie

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1969-2010	1970-2010	1969-2010	1970-2010
Berlin (Diplom)	0.55	0.54	0.08	0.08
Tübingen	0.62	0.62	0.25	0.24
Münster	0.67	0.68	0.17	0.17
Heidelberg (ab 1970)	0.67	0.67	0.23	0.23
Göttingen	0.70	0.71	0.15	0.15

Die Entwicklung der Noten an den einzelnen Hochschulen seit 1960 im Verlauf zeigt in Göttingen einen langen zyklischen Verlauf von ca. 30 Jahren ohne monotonen Trend. In Tübingen zeigt sich eine ähnliche lange Wellenbewegung mit niedrigerem Verlauf. In Münster sind zwei kürzere, recht stabile Zyklen zu erkennen, nur in Berlin und Heidelberg lassen sich Abwärtstrends erkennen. In Berlin ist eine deutliche Abwärtsbewegung sichtbar, die sich zu Beginn der 1980er in eine Aufwärtsphase umwandelt, der ein relativ stabiles Notenniveau folgt. In Heidelberg kann ein Abwärtstrend ab Ende der 1980er Jahre festgestellt werden, der im Gegensatz zu Göttingen nicht nur das Ausgangsniveau der zuvor durchlaufenen Aufwärtsphase erreicht, sondern weiter sinkt. Wird die Reihe um die Trendkomponente bereinigt, zeigen sich kürzere Schwankungen im Bereich von 10 Jahren.

Im Vergleich der letzten 10 Jahre der Reihe mit den ersten findet sich in Heidelberg durch dieses fortlaufende Sinken die stärkste Verbesserung, aber auch an den anderen Hochschulen findet sich mit Ausnahme von Tübingen ein niedrigeres Niveau im letzten Abschnitt der Reihe - trotz der Abwesenheit eines klaren Abwärtstrends, was aufgrund der zyklischen Struktur der Notenverläufe vor allem durch die forschungspragmatische Festlegung des Endzeitpunkts der Reihe bedingt sein dürfte. Die Zeitreihen einige Jahre weitergedacht, würde sich zumindest in Göttingen, Tübingen und Münster nach einigen Jahren auch im 10-Jahresmittel ein ausgeglichenes Niveau einstellen.

Tabelle 48: Kennzahlen - Notenentwicklung an den Hochschulen im Studiengang Soziologie 1969-2010

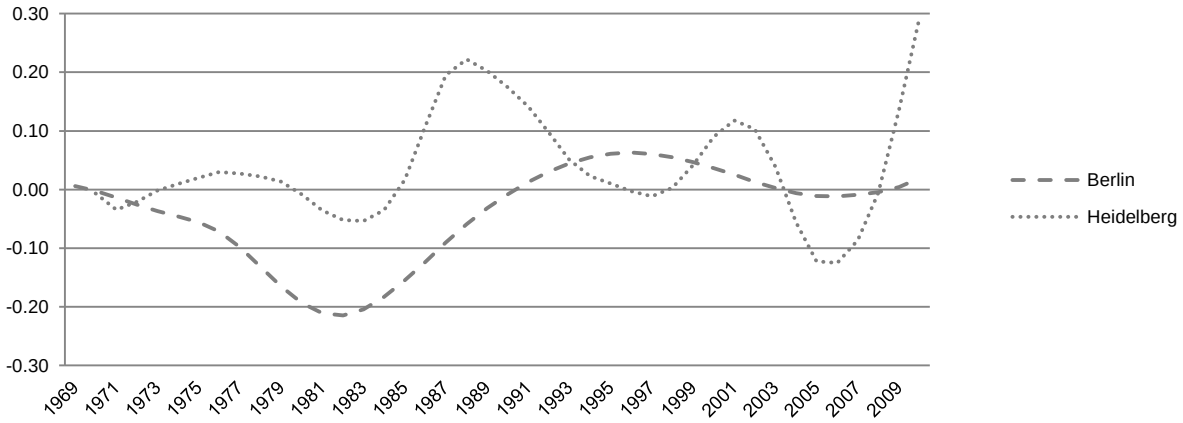
Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Heidelberg (ab 1970)	-0.31	-0.50	-0.013	-1.60 (1989-2007)	-0.089
Berlin (Diplom)	-0.19	-0.41	-0.010	-0.78 (1969-1982)	-0.060
Göttingen	-0.15	+0.08	+0.002	--	--
Münster	-0.11	+0.50	+0.012	--	--
Tübingen	+0.04	+0.17	+0.004	--	--

Tabelle 49: Kennzahlen - Notenentwicklung an den Hochschulen im Studiengang Soziologie 1970-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Heidelberg	-0.31	-0.50	-0.013	-1.60 (1989-2007)	-0.089
Göttingen	-0.19	+0.16	+0.004	--	--
Münster	-0.15	-0.17	-0.004	--	--
Berlin (Diplom)	-0.13	-0.17	-0.004	-0.64 (1973-1982)	-0.071
Tübingen	+0.01	-0.08	-0.002	--	--

Die Notenentwicklung an den einzelnen Magisterhochschulen stimmt somit auch weitestgehend mit dem Verlauf der Noten auf Studiengangebene überein. An drei der vier Universitäten verläuft das Notenniveau in einer langen Welle, die sich auch über alle Prüflinge gemittelt zeigt. Das Münsteraner Notenniveau erreicht ungefähr zum Zeitpunkt des Peaks dieser Wellen auch den Höhepunkt eines kürzeren Zyklus. Da die Noten in Heidelberg, Göttingen und Tübingen oberhalb des Studiengang-durchschnitts ähnlich verlaufen und mit Münster innerhalb der Hochschulen mit Magisterabschluss nur ein Gegenpol unterhalb dieses Durchschnitts existiert, der auch noch einen ähnlichen Verlauf nimmt, verändert sich das Verhältnis der Notenniveaus innerhalb der Magisterstudiengänge kaum. Lediglich gegen Ende der Reihe findet eine Angleichung der drei lange Zeit schlechteren mit dem meist besseren Münsteraner Niveau statt. Auch das Berliner Niveau passt sich nach anfänglicher Entfernung in Richtung besserer Noten im Zeitverlauf zunehmend den sich verbessernden Magister-noten an. Ausgeprägte Plateauphasen lassen sich nur in Berlin (1997-2010) finden. Wie in Psychologie ist auch hier dieses Mal nicht das Aufeinandertreffen von zyklischer Aufwärtsbewegung und schwachem Abwärtstrend der Grund, sondern schlicht mangelnde Dynamik in der Notengebung.

Abbildung 115: : Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.3-LOWESS 0.9)



Germanistik Magister[78]

Wie die grafische Darstellung erkennen lässt, wird die Spannweite zwischen bestem und schlechtestem Hochschuldurchschnitt in Germanistik zunehmend geringer. Beträgt sie bis zu Beginn der 1980er noch ca. 0.8 Noten, verringert sich diese Differenz bis unter 0.3 Noten. In Berlin werden über die meiste Zeit die besten Noten vergeben, in Göttingen bzw. in zwei Phasen in Münster die schlechtes-

[78] Anmerkung zur Datenbasis: Da die Fallzahlen in Germanistik vor 1970 sehr niedrig sind (hochschulübergreifend maximal bei n=13 in 1969), wird in der hochschulspezifischen Analyse nur der Zeitraum ab 1970 betrachtet. Auch in diesem Zeitraum sollte jedoch beachtet werden, dass, je nach Hochschule teilweise noch bis zu Beginn der 1980er Jahre Fallzahlen von n<10 pro Jahr vorliegen. Vor allem an der Universität Saarbrücken können die Noten erst ab dem Jahr 1983 als aussagekräftig gelten, dort beläuft sich die Fallzahl in den Jahren 1970-1982 in sechs Jahren auf n=1, in 4 Jahren auf n=2, in 2 Jahren auf n=3 und in einem Jahr auf n=4. Es sind im Zeitraum von 1970 bis 2010 alle Hochschulen im sample enthalten, weshalb sich eine Aufteilung in zwei Zeiträume in diesem Fall erübrigt.

ten. Das Notenniveau in Tübingen, Heidelberg und Saarbrücken bewegt sich zwischen diesen Polen und ab Beginn der 1980er nahe am Durchschnittsniveau aller Prüflinge.

Abbildung 116: Durchschnittliche Abschlussnoten an den Hochschulen in Germanistik Magister - Zeitverlauf (LOWESS 0.3)

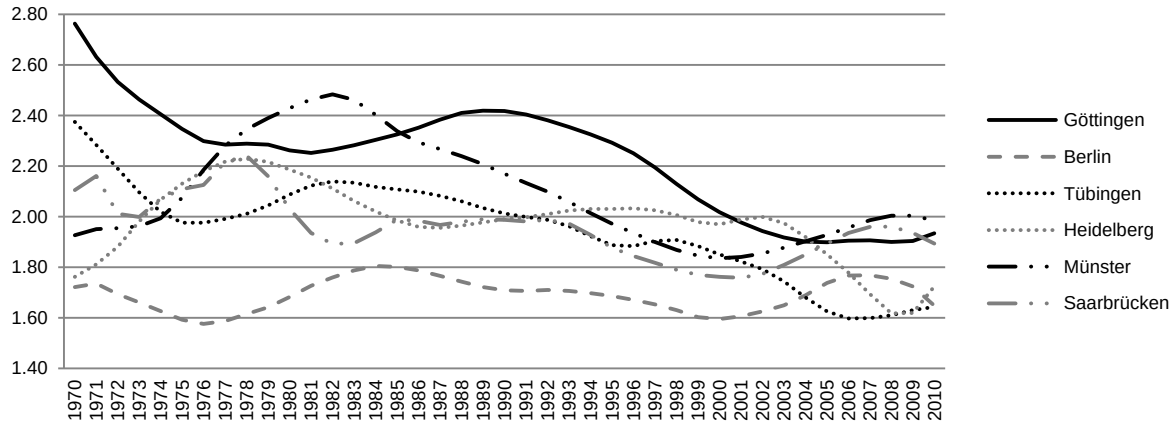
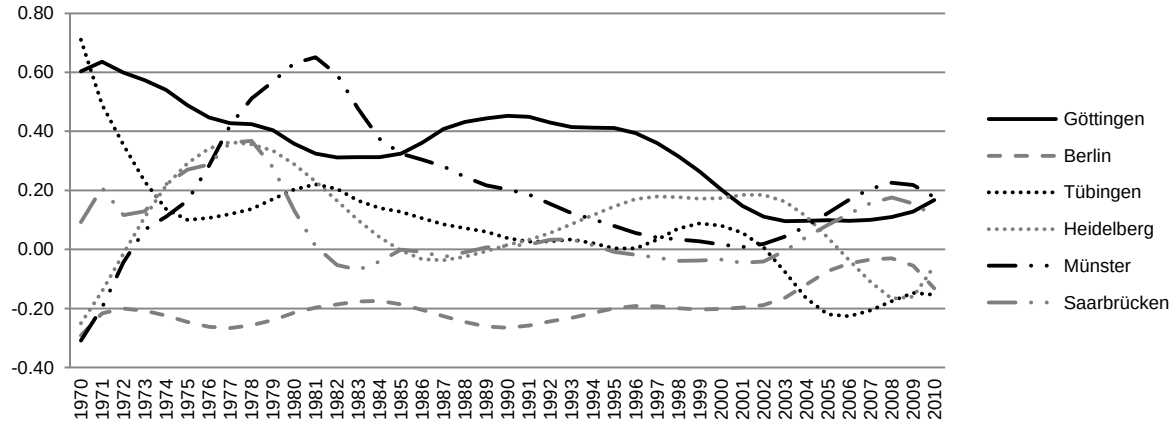


Abbildung 117: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Im Mittel über den gesamten Zeitraum sind die Noten der Göttinger Prüflinge durchschnittlich 0.34 Noten schlechter als die aller Prüflinge, in Berlin sind sie durchschnittlich 0.19 Noten besser. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt gemittelt bei R=0.85 und damit leicht höher als die grafische Aufbereitung der geglätteten Daten es vermuten lassen - nach VWL ist dies der zweithöchste Wert für alle Studiengänge in der Stichprobe. Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen über den gesamten Zeitraum gesehen 1983(R=0.27), am größten 1972 (R=2.14). In den 1980ern ist der Tiefpunkt der Entwicklung erreicht, tendenziell nimmt die Spannweite jedoch auch anschließend noch weiter ab, wenn auch nur geringfügig. Eine über den gesamten Zeitraum relativ konstante Notenhierarchie existiert in Germanistik nicht, es lassen sich lediglich einige relativ konstante Einstufungen ermitteln: In Berlin werden bis 2005 immer die besten, in Göttingen meist die schlechtesten Noten (jeweils zweimal Rang 4 und Rang 5, viermal Rang 6 von 6 in acht 5-Jahres-Rangfolgen) vergeben. Heidelberg nimmt von 1971 bis 1995 immer Rang 3 ein, Münster von 1971-1995 mit Ausnahme eines 5-Jahres-Wertes (1981-1985: Rang 6) immer Rang 5.

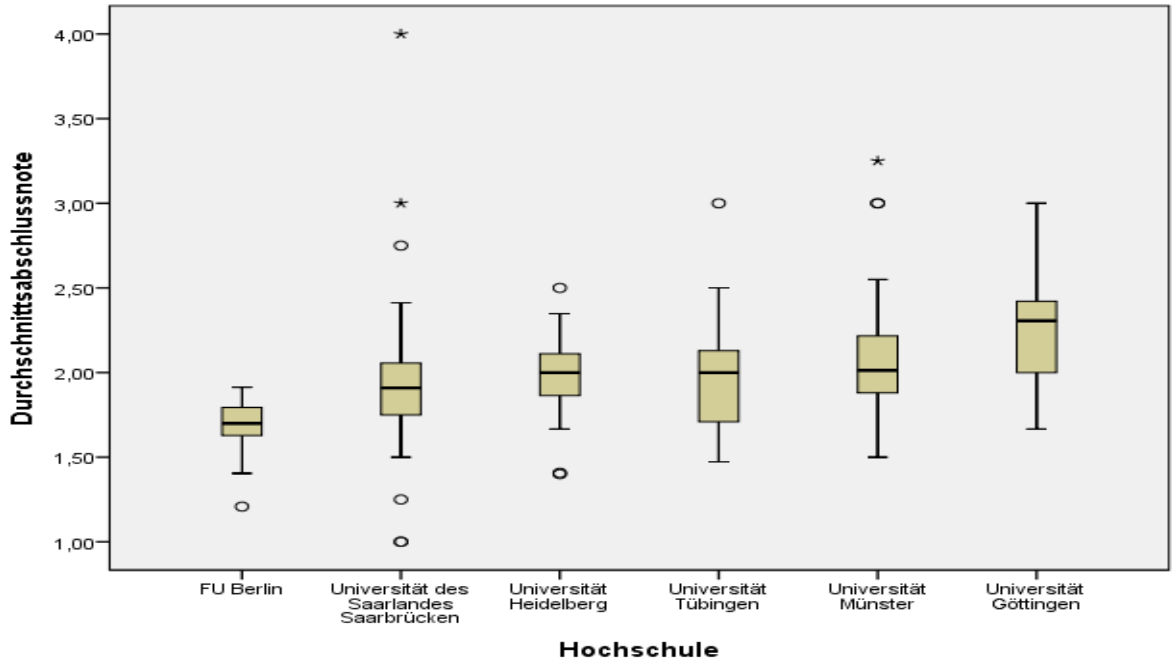
Tabelle 50: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

	1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	BER	1.68	BER	1.59	BER	1.80	BER	1.75	BER	1.71	BER	1.63	BER	1.63	HD	1.58
2	SB	1.85	TÜ	2.04	SB	1.82	SB	1.92	TÜ	1.96	SB	1.77	TÜ	1.72	TÜ	1.61
3	HD	1.96	HD	2.24	HD	2.07	HD	1.97	HD	2.03	TÜ	1.86	SB	1.79	BER	1.76
4	TÜ	2.01	GÖ	2.29	TÜ	2.11	TÜ	2.08	SB	2.04	MS	1.86	MS	1.84	GÖ	1.88
5	MS	2.03	MS	2.33	GÖ	2.28	MS	2.20	MS	2.06	HD	2.02	GÖ	1.90	SB	1.96
6	GÖ	2.45	SB	2.49	MS	2.41	GÖ	2.43	GÖ	2.33	GÖ	2.18	HD	2.01	MS	2.01

Die einfaktorielle ANOVA bringt signifikante Unterschiede zwischen mindestens zwei der Hochschulen in 34 der 41 Jahre seit 1970 hervor. Kruskal-Wallis- und Games-Howell-Test geben 33 von 41 Jahren als signifikant aus. Es unterscheiden sich die Noten in Heidelberg und Berlin am häufigsten signifikant voneinander - in 19 von 41 Jahren, bei durchschnittlich 48 bzw. 104 Prüflingen pro Jahr im Vergleichszeitraum. Ein ähnliches Verhältnis besteht zwischen Göttingen und Berlin (16 von 41 Jahren bei n=30 bzw. n=104). Bis auf den kurzen Zeitraum von 2001 bis 2005, in dem die Differenzen zwischen Tübingen und Heidelberg durchgehend Signifikanz aufweisen, sind drei oder mehr aufeinanderfolgende Jahre mit signifikanten Notenunterschieden nur in Paarvergleichen mit Berliner Beteiligung, und zwar für Göttingen (1981-1998), Heidelberg (1988-1998) und Münster (1988-1994), zu finden.

Über den gesamten Zeitraum gemittelt liegt die Distanz (die Betragsfunktion der Differenz zwischen den Notenniveaus) im Bereich von 0.27 (zwischen Tübingen und Heidelberg) bis 0.54 (zwischen Göttingen und Berlin) Noten. Für 11 der 15 Paarvergleiche liegt dieser Wert im Bereich von 0.30 bis 0.41. Die größte prozentuale Differenz zwischen den Notenniveaus lässt sich über die Jahre gemittelt zwischen Berlin und Göttingen feststellen. Das Notenniveau in Berlin erreicht durchschnittlich 77.2% des Göttinger Niveaus. Mit Ausnahme der guten Berliner Noten decken die mittleren 50% der seit 1970 vergebenen Noten an allen Hochschulen einen ähnlichen Bereich der Notenskala ab, wie die Überschneidungen der Boxen in Abbildung 118 zeigen. Die Berliner Noten sind dabei nicht nur niedrig, sondern auch sehr kompakt. Die Spannweite liegt selbst unter Berücksichtigung des Ausreißers noch ca. 0.4 Noten unter der nächstgrößeren in Heidelberg.

Abbildung 118: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1970-2010



Über die Jahre gemittelt zeigt sich in den Individualnoten eine enorm instabile Streuung in Saarbrücken und eine im Vergleich zu den anderen Hochschulen konstante Streuung in Berlin und Heidelberg. Göttingen, Münster und Tübingen liegen zwischen diesen Werten, allerdings ist auch hier die Streuung der Streuung im Zeitverlauf im Vergleich zu den Werten in anderen Studiengängen im oberen Bereich. Die Streuung selbst ist in Saarbrücken im Durchschnitt ebenfalls viel größer als an den anderen Standorten, die aber auch unter sich eine große Spannweite an Werten aufweisen, vor allem durch die vergleichsweise geringe durchschnittliche Standardabweichung in Berlin - die im Vergleich zwischen den Studiengängen allerdings immer noch im oberen Bereich der Werte liegt.

Tabelle 51: Streuung der Noten an den Hochschulen im Magisterstudiengang Germanistik 1970-2010

Hochschule	Mittlere Standardabweichung	Standardabweichung der Standardabweichung
Berlin	0.67	0.08
Heidelberg	0.71	0.10
Tübingen	0.71	0.17
Göttingen	0.76	0.17
Münster	0.77	0.15
Saarbrücken	0.87	0.31

Im Zeitverlauf lässt sich in Göttingen und Tübingen ein deutlicher Abwärtstrend erkennen, der in Göttingen auf einem höheren Niveau verläuft, aber zu einer ähnlichen Verbesserung in den 10-Jahresmittelwerten der 2000er gegenüber den 1970er Jahren führt und dort eine zunehmende Angleichung an die Niveaus der anderen Hochschulen bewirkt.

Die Trendbereinigung (Abb.119) lässt auch hier Zyklen, in Tübingen ca. 15-20, in Göttingen ca. 20 Jahre lang, hervortreten. In Münster bewegen sich die Noten in einer ca. 25-jährigen Wellenbewegung von Mitte der 1970er bis Ende der 1990er, die anschließend wieder eine Aufwärtsbewegung auf das ursprüngliche Niveau vollzieht, was möglicherweise den neuen Beginn eine derartigen Welle darstellt. In Berlin, Heidelberg und Saarbrücken sind etwas kürzere zyklische Bewegungen von ca. 20

Jahren zu erkennen, das niedrige Notenniveau in Berlin ist über die Zeit am stabilsten. In Heidelberg liegt die Durchschnittnote 2010 etwas unter dem Anfangswert, was allerdings an einer kurzen Abwärtsbewegung gegen Ende der Reihe liegt, die über die meiste Zeit relativ stabil verläuft und keinen echten Abwärtstrend aufweist.

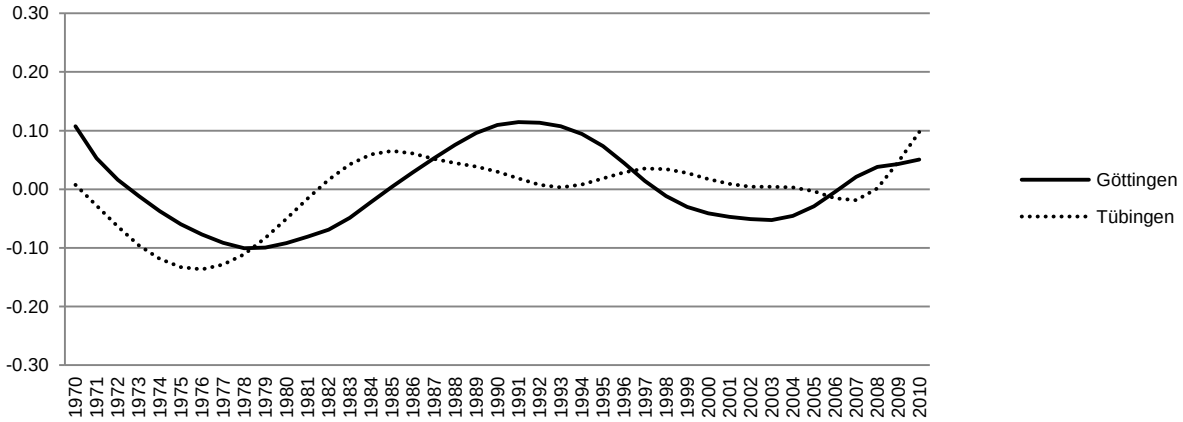
Auch die Differenz zwischen letztem und erstem 10-Jahres-Mittel impliziert eine langfristige Verbesserung, nicht nur in Heidelberg, sondern auch in Münster und Saarbrücken, was (wie in Soziologie) möglicherweise eher mit dem Ende der Datenreihen zu einem Zeitpunkt nach einer Phase niedriger Noten im zyklischen Gesamtbild zusammenhängt als mit einer tatsächlichen langfristigen Verbesserung. Das hier gebildete Maß der Verbesserung als Differenz zweier über 10 Jahre gemittelter Notenniveaus fungiert zwar als Schutz vor ausreißerbedingten Fehlinterpretationen, wenn ein Abwärtstrend vorliegt - für die Erfassung zyklischer Verläufe ist es offenbar nicht geeignet. In Saarbrücken geht der zyklische Verlauf scheinbar mit einem schwachen Abwärtstrend ab Ende der 1970er und einer leichten Verbesserung im Niveau 2010 gegenüber dem Ausgangsniveau einher. Wird aber nur der Zeitraum ab 1983 betrachtet, ab dem Fallzahlen n≥9 vorliegen, ist eine tatsächliche Verbesserungstendenz nur noch über einen Zeitraum von 11 Jahren (1991-2002) zu erkennen. Diese Periode stellt, wie die grafische Darstellung zeigt, eher die Abwärtsbewegung eines Zyklus dar, als einen Abwärtstrend.

Tabelle 52: Kennzahlen - Notenentwicklung an den Hochschulen im Magisterstudiengang Germanistik 1970-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Göttingen	-0.49	-0.33	-0.008	-1.31 (1971-2002)	-0.042
Tübingen	-0.45	-1.38	-0.035	-1.53 (1970-2005)	-0.044
Saarbrücken	-0.26	-0.21	-0.005	--	--
Heidelberg	-0.26	-0.05	-0.001	--	--
Münster	-0.12	+0.18	+0.004	--	--
Berlin	+0.07	+0.04	+0.001	--	--

An vier Hochschulen ist der gleiche zyklische Verlauf der Noten wie auf Studiengangebene erkennbar - ohne erkennbaren Abwärtstrend. In Göttingen und Tübingen werden die Noten entgegen dem Studiengangverlauf allerdings in zyklischen Bewegungen kontinuierlich besser, was eine Abweichung vom durchschnittlichen Verlaufsmuster in Germanistik Magister darstellt. Zusammen mit der langen Wellenbewegung in Münster sind die Noten an diesen beiden Hochschulen für die meiste Bewegung im Studiengang verantwortlich. Im Laufe der Abwärtsbewegungen findet eine Annäherung an das Niveau der anderen Hochschulen statt, so dass sich die Noten aller Hochschulen im Zeitverlauf zunehmend angleichen.

Abbildung 119: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.4-LOWESS 0.9)



Deutsch Lehramt[79]

Im Lehramtsstudiengang Deutsch zeigt sich zu Beginn der Zeitreihen eine Differenz von ca. 0.3 zwischen besten und schlechtesten Noten, die bis auf unter 0.2 in den 1970ern sinkt, bevor sie im Laufe der nächsten 10 Jahre auf 0.4 ansteigt, Anfang der 1990er wieder auf 0.2 sinkt und schließlich auf über 0.8 Punkte in den 2000ern ansteigt. Der Grund für diesen enormen Anstieg in den letzten Werten ist ein starker Abfall der Noten in Karlsruhe bei gleichzeitiger Verschlechterung der Berliner Noten. Ohne die Karlsruher Noten, die ab Mitte der 1970er Jahre mit Abstand das beste Durchschnittsniveau bilden, würde die Spannweite im Zeitverlauf nicht viel höher als bis 0.5 steigen - und auch dieser Wert würde nur durch den Anstieg der Berliner Werte erreicht. Über die meiste Zeit liegen alle Hochschulen außer Karlsruhe in ihrem durchschnittlichen Notenniveau relativ nahe beieinander. Die grafische Darstellung lässt keine einzelne Hochschule als dauerhafte obere Begrenzung der Spannweite erkennen. Berlin nimmt diese Position ab Mitte der 1970er Jahre am häufigsten ein, wird aber gegen Ende der 1980er - bis 1997, der letzte Zeitpunkt, für den Tübinger Werte vorliegen - abgelöst von Tübingen.

[79] Anmerkungen zur Datenbasis: In Deutsch Lehramt stehen Daten für mindestens zwei Hochschulen ab 1963 zur Verfügung. In der Karlsruher Reihe existiert eine Lücke von 1998-2004 (für diese Jahre liegen in der Prüfungsstatistik keine Informationen vor). Für Tübingen sind nach 1997 keine Informationen in der Prüfungsstatistik enthalten. In Braunschweig sind zu Beginn der Zeitreihe zwei Datenpunkte mit geringer Fallzahl (n<4) entfernt worden.
Die Durchschnittsnoten in Karlsruhe, Berlin und Tübingen enthalten die Noten aller im jeweiligen Landesprüfungsamt bestandenen Prüfungen, welche anhand der Zeugnisse nicht mehr eindeutig einzelnen Hochschulen zuzuordnen waren. So sind in den Karlsruher Noten auch die Absolvent*innen aus Mannheim und Heidelberg enthalten, in den Tübinger Noten die aus Ulm und in den Berliner Noten die, aller an Berliner Hochschulen abgelegten Lehramtsprüfungen.

Abbildung 120: Durchschnittliche Abschlussnoten an den Hochschulen in Deutsch Lehramt - Zeitverlauf (LOWESS 0.3)

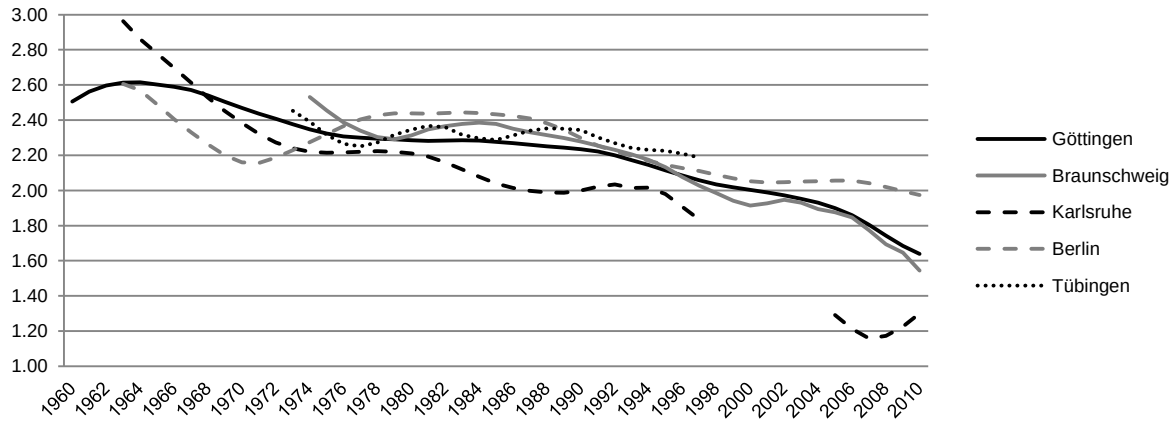
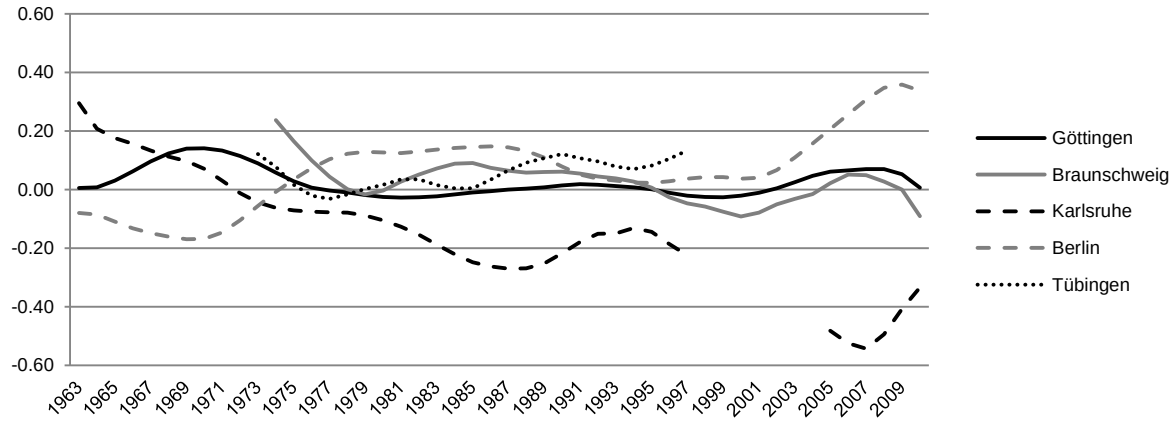


Abbildung 121: Differenz zwischen Hochschuldurchschnitt und Studiengangdurchschnitt (LOWESS 0.3)



Im Mittel über den gesamten Zeitraum sind die Noten der Berliner Prüflinge durchschnittlich 0.06 Noten schlechter als die aller Prüflinge, in Karlsruhe sind sie durchschnittlich 0.14 Noten besser. Wird nur der Zeitraum betrachtet, in dem für alle Hochschulen im sample Werte vorliegen (1973-1997), liegen die Berliner Noten 0.09 Noten über dem Durchschnitt, die Karlsruher 0.15 Noten darunter. Die Spannweite zwischen bestem und schlechtestem Notendurchschnitt liegt über alle Jahre seit 1963 gemittelt bei R=0.40 bzw. bei R=0.35 zwischen 1973-1997. Am geringsten ist die Differenz zwischen der Hochschule mit den besten und der mit den schlechtesten Ergebnissen über den gesamten Zeitraum gesehen 1965 (R=0.03) bzw. 1994 (R=0.17) für den Zeitraum in dem die Stichprobe komplett ist. Am größten ist diese Differenz 2005 (R=1.07) bzw. 1986 (R=0.56). Die drei letzten Werte spiegeln die Gesamtentwicklung wider: Die Spannweite steigt nach einer anfänglichen Abnahme von ca. 1980 an bis 1986, bevor sie bis 1994 wieder absinkt, um zu Beginn der 2000er wieder sprunghaft anzusteigen.

Eine konstante Notenhierarchie existiert nur in den Jahren von 1981-1990: Hier sind die Noten über 5 Jahre gemittelt in Karlsruhe am besten, es folgen Göttingen, dann Tübingen, Braunschweig und schließlich Berlin. Seit 1976 sind die Noten in Karlsruhe für den Zeitraum, für den dort Werte vorliegen, immer am besten.

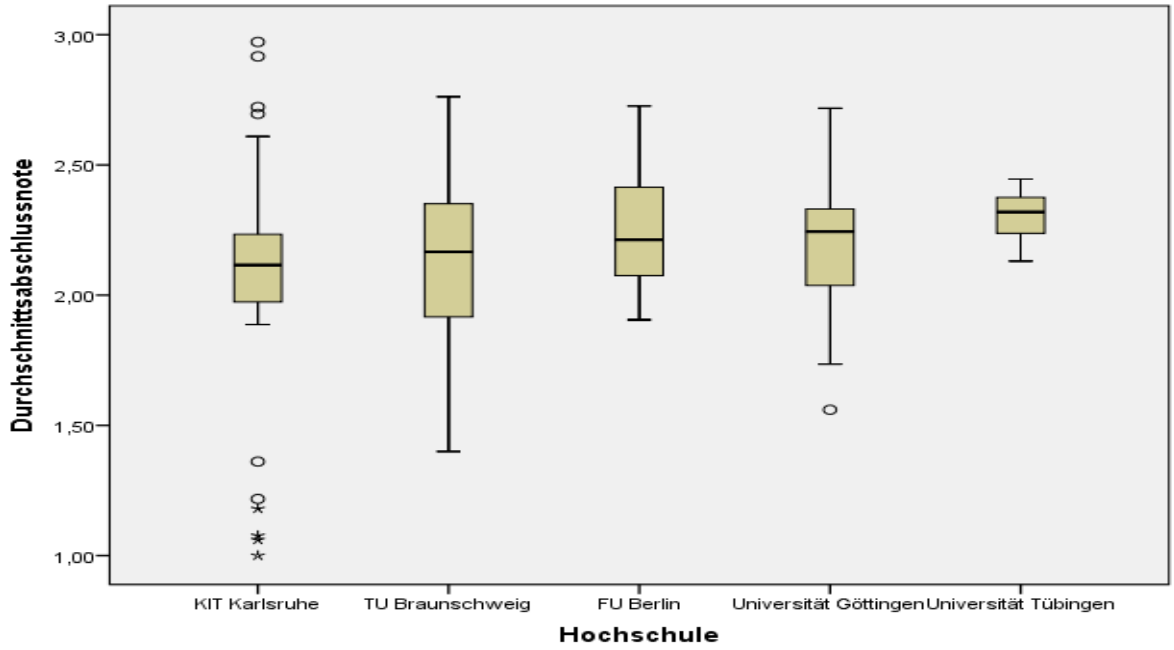
Tabelle 53: Rangfolge von Fünfjahresdurchschnitten der auf Hochschulebene gemittelten Durchschnittsnoten

	1961-1965		1966-1970		1971-1975		1976-1980		1981-1985		1986-1990		1991-1995		1996-2000		2001-2005		2006-2010	
1	BER	2.60	BER	2.22	BER	2.19	KA	2.23	KA	2.12	KA	1.99	KA	2.04	GÖ	1.96	KA	1.08	KA	1.16
2	GÖ	2.68	GÖ	2.53	KA	2.22	BS	2.26	GÖ	2.26	GÖ	2.28	BER	2.14	KA	1.97	BS	1.96	BS	1.66
3	KA	2.86	KA	2.54	GÖ	2.38	TÜ	2.27	TÜ	2.32	TÜ	2.34	BS	2.18	BS	2.01	GÖ	2.03	GÖ	1.73
4					TÜ	2.40	GÖ	2.28	BS	2.38	BS	2.35	GÖ	2.21	BER	2.12				
5					BS	2.55	BER	2.45	BER	2.42	BER	2.43	TÜ	2.26	TÜ	2.20	BER	2.05	BER	2.01

Die einfaktorielle ANOVA weist in 30 von 48 Jahren Signifikanz auf, Kruskal-Wallis und Games-Howell Test geben mit 31 signifikanten Werten beide ein Jahr mehr aus. Es unterscheiden sich die Noten in Karlsruhe und Berlin am häufigsten signifikant voneinander - in 19 von 41 Jahren, bei durchschnittlich 110 bzw. 103 Prüflingen pro Jahr im Vergleichszeitraum. Ein ähnliches Verhältnis besteht zwischen Karlsruhe und Tübingen (sieben von 25 Jahre bei n=137 bzw. n=133). Drei Jahre oder länger in Folge sind die Differenzen im Notenniveau nur zwischen Karlsruhe und den anderen vier Standorten signifikant. Gegenüber Göttingen dauert diese Phase nur von 1986-1988 an, gegenüber Braunschweig (2005-2008) und Tübingen (1985-1988) sind es ebenfalls nur kurze Zeiträume. Einzig zwischen Karlsruhe und Berlin gibt es zwei Phasen (1980-1988 und 2005-2010), in denen die signifikanten Differenzen etwas länger andauern.

Die Abstände zwischen den Noten der einzelnen Hochschulen sind recht ähnlich: Die Distanz (die Betragsfunktion der Differenz zwischen den Notenniveaus) liegt im Durchschnitt über alle Jahre im Bereich von 0.10 (zwischen Göttingen und Tübingen) bis 0.35 (zwischen Karlsruhe und Berlin) Noten. Für sieben der 10 Paarvergleiche liegt dieser Wert im Bereich von 0.14 bis 0.22. Über die Jahre gemittelt besteht die größte prozentuale Differenz zwischen den Notenniveaus zwischen Karlsruhe und Braunschweig. Das Notenniveau in Karlsruhe liegt durchschnittlich bei 86.7% des Braunschweiger Niveaus. Für den Zeitraum mit allen Hochschulen im sample liegen die Karlsruher Noten mit 89.4% des Berliner Niveaus am weitesten von einer anderen Hochschule entfernt. Die Karlsruher Noten zeichnen sich in ihrer Verteilung über den gesamten Zeitraum als ziemlich ausreißerbelastet aus, die Spannweite der Noten liegt unter Berücksichtigung aller Werte bei zwei ganzen Noten. Auch Braunschweig weist mit R=1.5 Noten Spannweite - ohne dabei Ausreißer aufzuweisen - eine große Bandbreite an vergebenen Noten im betrachteten Zeitraum auf, während Tübingen im kurzen Zeitraum für den dort Daten vorliegen gerade einmal 31% einer Note an Spannweite produziert. An den anderen Hochschulen ist damit gerade einmal der Interquartilsabstand erfasst, in Braunschweig ist dieser sogar noch mehr als eine Zehntelnote größer. Trotz dieser Differenzen liegen die Mediane aller fünf Hochschulen vergleichsweise nahe beieinander.

Abbildung 122: Jährlich gemittelte Abschlussnoten in ihrer Verteilung über die Jahre 1963-2010



Die Noten streuen über die Zeit am stabilsten in Tübingen. Die dortige Verteilung der Noten um den Mittelwert ist im Zeitverlauf deutlich homogener als an den anderen Hochschulen, von denen Braunschweig und vor allem Karlsruhe wesentlich höhere Streuungen der Streuung über die Zeit aufweisen. Im kurzen Vergleichszeitraum von 1974 bis 1997 gleichen sich die Werte in Karlsruhe, Göttingen und Berlin allerdings an den Tübinger Wert an. Die Stärke der durchschnittlichen Streuung unterscheidet sich im Gesamtzeitraum weniger stark. Die geringe Differenz zu Tübingen verringert sich im kürzeren Zeitraum in Karlsruhe und Braunschweig durch eine Erhöhung der Werte dort, während die Werte in Göttingen und Berlin kleiner werden, was die Heterogenität für diesen Zeitraum leicht erhöht. Gegenüber dem Magisterstudiengang Germanistik finden sich leicht niedrigere Streuungswerte - gegenüber den Werten im mathematischen Lehramtsstudiengang sind sie noch niedriger.

Tabelle 54: Streuung der Noten an den Hochschulen im Lehramtsstudiengang Deutsch

Hochschule	Mittlere Standardabweichung		Standardabweichung der Standardabweichung	
	1963-2010	1974-1997	1963-2010	1974-1997
Karlsruhe (1963-1997, ab 2005)	0.68	0.73	0.19	0.04
Göttingen	0.68	0.67	0.07	0.05
Berlin	0.70	0.68	0.09	0.05
Braunschweig (ab 1974)	0.70	0.73	0.13	0.11
Tübingen (1973-1997)	0.73	0.73	0.05	0.05

Im Zeitverlauf zeigt sich neben der starken Verbesserung in Karlsruhe auch an allen anderen Hochschulen, für die über einen längeren Zeitraum Daten vorliegen, ein deutlicher Abwärtstrend. Besonders in Karlsruhe ist eine sehr starke Abwärtsdynamik zu beobachten, die in diesem Ausmaß - 1.4 Noten Verbesserung im 10-Jahres-Mittelwertvergleich - sonst nur in VWL auftritt. In Göttingen, Braunschweig und Berlin hingegen nimmt die Verbesserung, in dieser Reihenfolge abnehmend, Ausmaße an, die an den Hochschulen auch in den anderen Studiengängen, in denen langfristige Ab-

wärtsbewegungen zu beobachten sind, auftreten. Im stärksten Trendbereich, der abgesehen von der TU Braunschweig, für die erst 1974 Werte vorliegen, überall gleichzeitig eintritt und nur in Berlin einige Jahre früher endet, erreichen die Hochschulen jährliche Verbesserungsstärken zwischen 0.022 Noten in Berlin und 0.039 Noten in Braunschweig. Nur in Tübingen kann für den kurzen Zeitraum von 1973 bis 1997 keine eindeutige Abwärtsbewegung festgestellt werden. Allerdings fallen in dem Zeitraum, für den in Tübingen Daten vorliegen, auch die Verbesserungen in den anderen Studiengängen geringer aus. Die absolute Veränderung der Noten in diesem Zeitraum entspricht in Stärke und Höhe allen anderen Hochschulen, was ein Hinweis darauf ist, dass ein eindeutiger Abwärtstrend in Tübingen nur aufgrund der mangelnden Datenlage nicht erfasst werden konnte. Alle Notenbewegungen weisen auch hier Zyklen von 10-20 Jahren Länge auf, wie die Trendbereinigung (Abb.123) zeigt.

Tabelle 55: Kennzahlen - Notenentwicklung an den Hochschulen im Lehramtsstudiengang Deutsch 1963-2010

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Karlsruhe (1963-1997, ab 2005)	-1.42	-1.70	-0.021	-1.97 (1964-2007)	-0.031
Göttingen	-0.68	-1.11	-0.024	-1.16 (1965-2010)	-0.026
Braunschweig (ab 1974)	-0.55	-0.93	-0.026	-1.36 (1975-2010)	-0.039
Berlin	-0.27	-0.59	-0.013	-0.82 (1965-2002)	-0.022
Tübingen (1973-1997)	-0.06	-0.26	-0.011	--	--

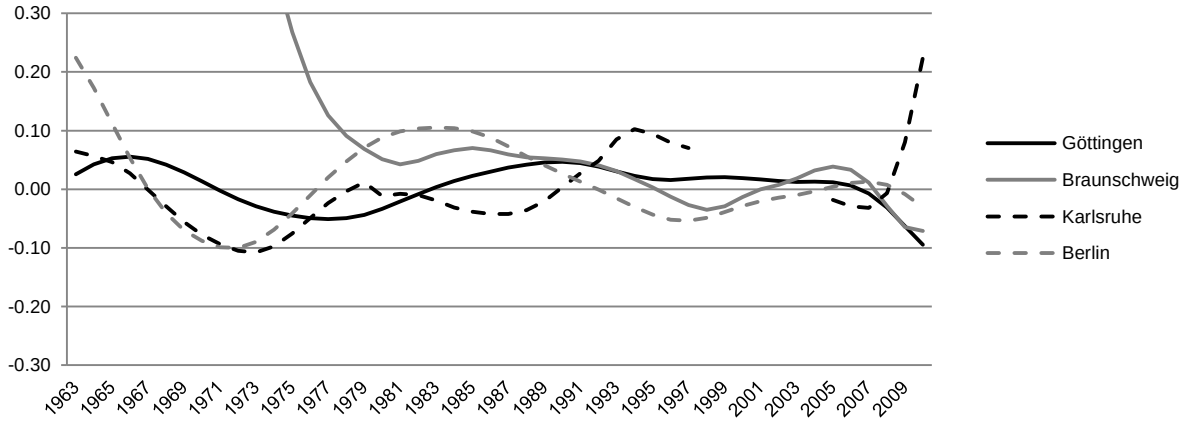
Tabelle 56: Kennzahlen - Notenentwicklung an den Hochschulen im Lehramtsstudiengang Deutsch 1974-1997

Hochschule	Letzte 10 minus erste 10 Jahre	Veränderung absolut	Veränderung /Jahr	Veränderung im stärksten Trendbereich	Veränderung /Jahr
Berlin	-0.20	-0.17	-0.007	-0.52 (1986-1997)	-0.047
Karlsruhe	-0.19	-0.19	-0.008	-0.43 (1979-1988)	-0.048
Braunschweig	-0.15	-0.19	-0.008	-0.71 (1975-1995)	-0.036
Göttingen	-0.11	-0.26	-0.011	-0.44 (1975-1997)	-0.020
Tübingen	-0.06	-0.23	-0.010	--	--

Der Verlauf der Noten auf Studiengangebene spiegelt sich damit auch im Lehramtsstudiengang Deutsch auf Hochschulebene wider. An vier der fünf Universitäten bzw. Landesprüfungsämter (Göttingen, Braunschweig, Karlsruhe, Berlin) verläuft das Absinken des Notenniveaus in den Wellen, die sich auch über alle Prüflinge gemittelt zeigen. Ausgeprägte Plateauphasen lassen sich in Tübingen (19985-1991), Berlin (1977-1985) und Göttingen (1978-1995) erkennen. Für letztere beiden ist diese Phase vermeintlicher Stabilität der Noten wiederum durch die zeitliche Parallelität von zyklischer Aufwärtsbewegung und Abwärtstrend zu erklären. Für die Tübinger Reihe lässt sich aufgrund der kurzen Dauer nicht sagen, ob die Noten durch mangelnde Dynamik oder das Aufeinandertreffen zweier gegensätzlicher Verlaufskomponenten in dieser Phase stabil sind.

Im Verhältnis der Notenniveaus an den Hochschulen verändert sich im Zeitverlauf kaum etwas, vielmehr ist eine parallele Entwicklung zu beobachten. Die Ausnahme bildet Karlsruhe, wo sich die Noten im Vergleich zu den anderen Hochschulen zunehmend stark verbessern und sich damit weiter von den Noten dort und vom Durchschnitt im gesamten Studiengang entfernen.

Abbildung 123: Trendbereinigte Zeitreihen der Hochschulen mit zyklischer Komponente (LOWESS 0.4-LOWESS 0.9)



Die hochschulspezifische Notengebung im Vergleich zwischen den Studiengängen

Tabelle 57 fasst die wichtigsten Kennzahlen der Unterschiede im Notenniveau der studiengangintern-nen Hochschulvergleiche zusammen: Die über alle Jahre gemittelte Spannweite zwischen bester und schlechtester Abschlussnote im Studiengang (Spalte 3), die Hochschule, die am häufigsten den besten bzw. den schlechtesten Notendurchschnitt aufweist (Spalten 4 und 5, berechnet als Anteil der Jahre, in denen im Beobachtungszeitraum die beste/schlechteste Durchschnittsnote vorliegt an der Gesamtzahl der Jahre, in denen für die Hochschule Werte vorliegen), der über den gesamten Zeitraum gemittelt niedrigste prozentuale Anteil des Notenniveaus einer Hochschule an dem einer anderen (Spalte 6), die über alle Paarvergleiche der Post-Hoc Tests gemittelte durchschnittliche Distanz (Betragsfunktion der Differenz im Notenniveau) zwischen den Notenniveaus der Hochschulen (Spalte 7) und den Bereich und die Spannweite der über alle Jahre gemittelten Standardabweichung der Durchschnittsnoten (Spalte 8) sowie Bereich und Spannweite der Standardabweichung dieser gemittelten Standardabweichung (Spalte 9).

Die Studiengänge sind in der Tabelle nach steigender durchschnittlicher Spannweite sortiert. Das homogenste Notenniveau findet sich über alle Jahre des Beobachtungszeitraums gemittelt im Lehramtsstudiengang Deutsch, das heterogenste in VWL - dort liegen beste und schlechteste Durchschnittsnoten mehr als doppelt so weit auseinander wie in Deutsch auf Lehramt. Es fällt auf, dass die Spannweite von Studiengang zu Studiengang in recht kleinen Schritten ansteigt, bevor zwischen Mathematik Lehramt und Chemie ein deutlicher Abstand auftritt. Beträgt die maximale Differenz zwischen bester und schlechtester Durchschnittsnote im mathematischen Lehramtsstudium in der Stichprobe im Durchschnitt über alle Jahre noch 0.60 Noten, sind es in Chemie schon 0.72 Noten. Dort liegen die Abschlussnoten an der Hochschule mit dem besten und der mit dem schlechtesten Notenniveau im Mittel also über eine Zehntelnote weiter auseinander. Werden die Studiengänge basierend auf der Differenz zwischen höchster und niedrigster Spannweite in zwei Gruppen aufgeteilt, von denen eine die Studiengänge umfasst, deren Werte unter der Hälfte dieser Differenz liegen,

und die andere, diejenigen, deren Werte darüber liegen, so liegt der Trennstrich genau dort - zwischen Mathematik Lehramt und Chemie. Werden noch die anderen Indikatoren für die Homogenität des Notenniveaus hinzugezogen, zeigt sich, dass die Studiengänge mit einer niedrigen durchschnittlichen Spannweite nicht weiter überraschend auch einen höheren durchschnittlichen Anteil des niedrigsten am höchsten Notenniveau aufweisen, aber auch dass dort die über alle Paarvergleiche gemittelte Distanz zwischen den Notenniveaus sowie die Spannweite der gemittelten Standardabweichung geringer ausfallen.

Das Distanzmaß, das als einziger dieser Werte nicht nur den maximalen Abstand zwischen bestem und schlechtestem Notenniveau erfasst, sondern die Abstände zwischen allen Hochschulen untereinander abbildet, belegt, dass die hier verwendeten Maße zur Erfassung der Homogenität der Notenniveaus innerhalb der Studiengänge geeignet sind. Die Korrelation zwischen der über alle Paarvergleiche gemittelten Distanz und der durchschnittlichen Spannweite zwischen bester und schlechtester Durchschnittsnote beträgt r=0.940 (p=0.000). Alle vier Maße untereinander weisen keinen geringeren Korrelationskoeffizienten als r=0.679 auf (wobei der Zusammenhang zwischen dem niedrigsten mittleren prozentualen Anteil des Notenniveaus zwischen zwei Hochschulen und den anderen drei Maßen natürlich negativ ist).

Grundsätzlich lassen sich also Studiengänge mit eher homogenen und Studiengänge mit eher heterogenem Notenniveau an den in der Stichprobe enthaltenen Hochschulen unterscheiden: Mathematik, Biologie, Psychologie, BWL, Deutsch Lehramt und Soziologie (wenn nur der Magisterstudiengang betrachtet wird) weisen ein eher einheitliches, Chemie, VWL und Germanistik ein eher uneinheitliches Notenniveau auf. Mathematik Lehramt muss als Grenzfall eingeordnet werden. Zwar ist die Spannweite dort als vergleichsweise niedrig einzuordnen und liegt nur 0.03 Differenzpunkte über der in Psychologie. Allerdings liegt die durchschnittliche mittlere Distanz hier schon im Überschneidungsbereich der Werte der Studiengänge mit eher homogenen und der mit eher heterogenem Notenniveau. Zudem liegt der niedrigste gemittelte prozentuale Anteil des Notenniveaus einer Hochschule an dem einer anderen fast genau in der Mitte zwischen den übrigen Studiengängen mit niedriger und denen mit hoher Spannweite, die Spannweite der gemittelten Standardabweichung sogar eindeutig im Wertebereich Letzterer.

Dass die nach Studiengang unterschiedliche Anzahl der in der Stichprobe enthaltenen Hochschulen einen Einfluss auf die Zuordnung hat, dürfte unwahrscheinlich sein, da in beiden Gruppen Studiengänge mit der maximalen Anzahl an Hochschulen als auch mit geringeren Zahlen vertreten sind. Aufgrund der Begrenzung des Notenniveaus auf den Wertebereich 1-4 und des im Vergleich zur Gesamtzahl deutscher Hochschulen niedrigen Stichprobenumfangs ist jedoch denkbar, dass sich die Wertebereiche der heterogenen Studiengänge bei einer Betrachtung aller Hochschulen ‚auffüllen', die der homogenen erweitern würden und die Aufteilung sich dadurch auflöst. Die Spannweiten, die sich

im Notenniveau über alle in den Wissenschaftsratsberichten enthaltenen Hochschulen berechnen lassen, weisen zwar ebenfalls Abstufungen auf, welche in ihrer Rangordnung aber nicht der ermittelten Reihenfolge entsprechen.

Werden nur die Spannweiten der Jahre 2000 und 2010 aus der Stichprobe mit den entsprechenden Werten des Wissenschaftsrats verglichen, zeigt sich nur für VWL eine Übereinstimmung in beiden Jahren, in Chemie und Biologie jeweils in einem Jahr. Alle anderen Werte liegen in der Stichprobe deutlich unter den Werten des Wissenschaftsrats, was die Aussagekraft der hier vorgenommenen Einteilung sehr stark einschränkt. Die Abstände zwischen den Notenniveaus der Hochschulen scheinen im Gegensatz zum mittleren Notenniveau der Studiengänge nicht die tatsächlichen Verhältnisse widerzuspiegeln. Die Unterscheidung in eher homogene und eher heterogene Notenniveaus kann dementsprechend nur als Orientierung innerhalb der Stichprobendaten dienen und sollte nicht als robustes Ergebnis verstanden werden.

In Bezug auf die Begrenzung der Stichprobe und damit für den Vergleich untereinander sind ebenfalls Auffälligkeiten in den Notenrangfolgen zu sehen: In Mathematik, Biologie, Psychologie und BWL werden in Berlin am häufigsten die besten Noten vergeben. Auch in Germanistik Magister und in der abschlussübergreifenden Betrachtung der Soziologienoten ist dies der Fall (Spalte 4).Und auch in VWL liegen die Berliner Noten mit 19 von 51 Jahren, in denen sie die besten sind, nicht weit hinter denen der Karlsruher Absolvent*innen, die in 25 von 48 Fällen den Spitzenplatz einnehmen. In Chemie hingegen liegt Berlin anteilsmäßig auf Platz 4, in den beiden Lehramtsstudiengängen sogar auf dem letzten Platz (Spalte 5). In Göttingen werden dagegen in den meisten Studiengängen am häufigsten die schlechtesten Noten an den erfassten Hochschulen vergeben - in Mathematik Diplom, VWL und in den Magisterstudiengängen, in BWL und Psychologie sind sie am zweithäufigsten die schlechtesten. In Mathematik Lehramt und Chemie hingegen sind die Göttinger Noten anteilsmäßig am häufigsten die besten. Auch in Karlsruhe finden sich beide Extreme: Am häufigsten beste Noten in Deutsch Lehramt und VWL, am häufigsten schlechteste Noten in Chemie und Biologie. Tübingen teilt sich den letzten Platz in Soziologie (beide Abschlüsse) mit Göttingen und nimmt den ersten in Psychologie ein, in Münster gibt es in Soziologie die meisten Bestnoten im Zeitverlauf, wenn nur die Magisterabschlüsse berücksichtigt werden, dafür die schlechtesten in BWL. Braunschweig und Heidelberg tauchen in diesen Rangfolgen nie ganz oben oder ganz unten auf, dort sind die Noten also in keinem Studiengang auffällig häufig besonders gut oder besonders schlecht.

Tabelle 57: Kennzahlen der Hochschulunterschiede nach Studiengängen

Studiengang / Betrachteter Zeitraum	Anzahl Hochschulen	Durchschnittliche Spannweite der Noten	Beste Noten am häufigsten in:	Schlechteste Noten am häufigsten in:	Anteil des besten am schlechtesten Notenniveau	Durchschnittliche mittlere Distanz über alle Paarvergleiche	Bereich / Spannweite der mittleren Standardabweichung	Bereich / Spannweite der Standardabweichung der mittleren Standardabweichung
Deutsch Lehramt 1963-2010	5	0.40	Karlsruhe (25/41)	Berlin (21/48)	86.7% (Karlsruhe-Berlin)	0.20	0.68-0.73 / 0.05	0.05-0.19 / 0.14
Biologie 1969-2010	7	0.42	Berlin (14/42)	Karlsruhe (12/28)	84.8% (Heidelberg-Karlsruhe)	0.17	0.52-0.57 / 0.05	0.04-0.13 / 0.09
BWL 1960-2010	4	0.49	Berlin (42/51)	Münster (31/51)	85.4% (Berlin-Münster)	0.25	0.57-0.65 / 0.08	0.05-0.07 / 0.02
Mathematik 1960-2010	7	0.52	Berlin (16/51)	Göttingen (21/51)	86.2 % (Heidelberg-Göttingen)	0.23	0.66-0.77 / 0.11	0.09-0.19 / 0.10
Soziologie (Magister) 1969-2010	4	0.55	Münster (21/42)	Göttingen (18/42)	88.5% (Münster-Tübingen)	0.31	0.62-0.70 / 0.08	0.15-0.25 / 0.10
Psychologie 1960-2010	6	0.57	Berlin (19/40)	Tübingen (17/51)	83.1% (Berlin-Göttingen)	0.25	0.51-0.62 / 0.11	0.08-0.15 / 0.07
Mathematik Lehramt 1961-2009	5	0.60	Göttingen (20/50) Tübingen (10/25)	Berlin (33/47)	80.4% (Tübingen-Berlin)	0.30	0.64-0.81 / 0.17	0.07-0.16 / 0.09
Chemie 1960-2010	7	0.72	Göttingen (27/51)	Karlsruhe (26/51)	71.9% (Göttingen-Karlsruhe)	0.31	0.51-0.76 / 0.25	0.06-0.15 / 0.09
Soziologie (+Diplom) 1969-2010	5	0.75	Berlin (27/42)	Göttingen (17/42) Tübingen (17/42)	77.9% (Berlin-Tübingen)	0.37	0.55-0.70 / 0.15	0.08-0.25 / 0.17
Germanistik 1970-2010	6	0.85	Berlin (24/41)	Göttingen (18/41)	77.2% (Berlin-Göttingen)	0.37	0.67-0.87 / 0.20	0.08-0.31 / 0.23
VWL 1960-2010	6	0.88	Karlsruhe (25/48)	Göttingen (23/48)	74.9% (Karlsruhe-Göttingen)	0.40	0.52-0.67 / 0.15	0.08-0.25 / 0.17

In den meisten Studiengängen überwiegt die Anzahl der Hochschulen, die den gleichen Notenverlauf aufweisen wie der Gesamtdurchschnitt (Tab.58). Lediglich in Biologie und Chemie ist das Verhältnis mit je vier Standorten mit langfristigem Abwärtstrend und drei ohne einigermaßen ausgeglichen. In Biologie stellen die vier Universitäten mit den sich verbessernden Noten bis 1980 einen deutlich größeren Anteil an Absolvent*innen als nach 1980 - von den Hochschulen ohne Verbesserung sind vor 1980 nur Daten aus Göttingen vorhanden - wodurch sich die Verbesserung auch auf Studiengangebene zeigt. Anschließend gleich sich das Verhältnis an Prüflingen zunehmend an, so dass die zyklischen Bewegungen der Noten an den drei Hochschulen ohne langfristige Verbesserung die Abwärtsbewegungen der Hochschulen mit Verbesserung abfedern. Deshalb zeigt sich auf Studiengangebene nur noch ein minimales Absinken der Noten in Biologie seit Mitte der 1970er Jahre mit gelegentlichen Ausreißern nach oben, gut sichtbar zu Beginn der 2000er Jahre. Dort erreichen die Noten an allen drei Hochschulen einen ihrer Peaks. Dass die zyklischen Bewegungen im restlichen Zeitverlauf auf Studiengangebene nicht weiter auffallen, liegt daran, dass sie im Vergleich zur Gesamtskala nur eine niedrige Spannweite aufweisen. In Chemie hingegen liegen die Absolvent*innenzahlen der Hochschulen mit Verbesserung zusammengerechnet den gesamten Zeitraum über denen der drei Einrichtungen ohne langfristigen Abwärtstrend - seit Beginn der 1970er sind sie ungefähr doppelt so hoch, vorher sogar noch höher. Daher ist es nicht weiter verwunderlich, dass sich auch auf Studiengangebene der Trend zur Verbesserung durchsetzt.

Die Hochschuldaten zeigen auch, dass eine zu kleine Stichprobe (wie zu erwarten) zu Verzerrungen führt: So ist der in Maschinenbau aufgezeigte zyklische Verlauf nur in Braunschweig zu finden - seine Entsprechung auf der höheren Aggregatebene beruht auf der Dominanz des dortigen Absolvent*innenanteils in genau dem Zeitraum, in dem in Karlsruhe eine Plateauphase die langfristige Verbesserung unterbricht. Ein charakteristischer Verlauf für den Studiengang Maschinenbau kann anhand der verfügbaren Daten demnach nicht bestimmt werden.

Im Vergleich der Studiengänge verläuft die langfristige Verbesserung der Hochschulen, an denen sie zu finden ist, in Biologie am einheitlichsten. Die Differenz im Notenniveau der letzten 10 Jahre mit dem der ersten 10 Jahre der Reihe reicht dort innerhalb der Hochschulen mit Abwärtstrend von -0.28 bis -0.16 Noten Verbesserung, was eine Spannweite von R=0.12 in diesem Verbesserungsmaß bedeutet. Am größten ist der Abstand zwischen niedrigster und höchster Verbesserung in Deutsch Lehramt, wo die Differenz 1.36 Noten beträgt. Auch in Psychologie und VWL ist noch über eine Note Differenz festzuhalten, in Mathematik Diplom und Lehramt, in Chemie und BWL sind die Werte deutlich niedriger. Im Vergleich zu beachten ist allerdings, dass die ersten und letzten 10 Jahre der einzelnen Zeitreihen weder innerhalb der Studiengänge noch zwischen diesen immer denselben Zeitraum abbilden. Die Spannweiten der Veränderung im stärksten Trendbereich entsprechen in den meisten Studiengängen in etwa dem Eindruck der Veränderung in den 10-Jahresmitteln. Nur in Mathematik liegen

die Spannweiten deutlich höher, im Diplom bei knapp einer ganzen Note, im Lehramt sogar bei 1.20 Noten. Auch bezüglich dieser Längsschnittkennzahlen sei jedoch noch einmal einschränkend auf die geringe Stichprobengröße und damit auf die mangelnde Übertragbarkeit von Homogenitäts- bzw. Heterogenitätsmaßen auf die Gesamtheit der Hochschulen in den einzelnen Studiengängen hingewiesen.

Tabelle 58: Die Notenentwicklung an den Hochschulen im Vergleich zum Studiengangtrend

Studiengang	Trend zur Verbesserung auf Studiengangebene	Hochschulen im sample	Hochschulen mit Trend zur Verbesserung	Hochschulen ohne Trend zur Verbesserung
VWL Diplom	ja	6	6	0
Mathematik Lehramt	ja	5*	5	0
Mathematik Diplom	ja	7	6	1
Psychologie Diplom	ja	6	5	1
BWL Diplom	ja	4/5**	4	0
Deutsch Lehramt	ja	5*	4	1
Biologie Diplom	ja	7	4	3
Chemie Diplom	ja	7	4	3
Jura 1. Staatsexamen	nein	10/11***	2	8
Germanistik Magister	nein	6	2	4
Soziologie Magister	nein	4	1	3
Maschinenbau Diplom	nein	2	1	1

*Die Lehramtsnoten sind in drei von fünf Fällen nur zu ganzen Landesprüfungsämtern zuzuordnen, umfassen dann mehrere Hochschulen. **Für Karlsruhe liegt nur eine sehr kurze Zeitreihe vor (1964-1981) ***Bundesländer statt Hochschulen, für Bremen liegt nur eine sehr kurze Zeitreihe vor (1990-2007)

Unabhängig von der mangelnden Übertragbarkeit der unterschiedlichen quantitativen Ausmaße der Differenzen im Notenniveau und dessen Entwicklung auf die Gesamtheit deutscher Hochschulen existieren zwischen den Hochschulen im sample innerhalb der Studiengänge erklärungsbedürftige Differenzen im Notenniveau und mehr oder weniger stabile Unterschiede in der langfristigen Entwicklung der Noten. So verlaufen auch die an den einzelnen Hochschulen zu findenden langfristigen Verbesserungen im gleichen Studiengang nicht unbedingt auf dem gleichen Niveau und in gleichem Ausmaß, setzen die Verbesserungen nicht überall gleichzeitig ein und verlaufen die Zyklen teils in unterschiedlichen Notenhöhen. Auch die Streuung der Noten fällt im selben Studiengang teils unterschiedlich stark aus. In einigen Studiengängen ist diese Heterogenität in Notenniveau und -entwicklung innerhalb der Stichprobe größer als in andern, wobei diese Unterschiede vermutlich auf den geringen Stichprobenumfang zurückzuführen ist, wie ein Vergleich der Daten mit denen des Wissenschaftsrats nahelegt.

Gemeinsamkeiten finden sich an den Hochschulen vor allem hinsichtlich der langfristigen Entwicklung: In der Regel sind auch an den einzelnen Hochschulen zyklische Bewegungen von 10-20 Jahren Länge zu finden, die entweder einen langfristigen Abwärtstrend begleiten oder sich relativ gleichmäßig um ein konstantes Notenniveau herum bewegen. Die Abwärtstrends weisen in den meisten Studiengängen mit langfristiger Verbesserung einen einigermaßen parallelen Zeitraum auf, in dem der Trend am stärksten ist - auch wenn vereinzelt Abweichungen beim Beginn oder Ende dieser Zeiträume zu beobachten sind.

Zusammenfassung der deskriptiven Ergebnisse

Die Abschlussnoten an deutschen Hochschulen verteilen sich je nach Studiengang in unterschiedlicher Weise über die vier Notenstufen ‚sehr gut' bis ‚ausreichend'. In VWL und den beiden Lehramtsstudiengängen wird die Breite der Notenskala am ehesten ausgenutzt, in Biologie, Psychologie und Maschinenbau schließt nur ein verschwindend geringer Anteil der Studierenden mit der Note 4 ab. Es zeigen sich nicht nur fach- sondern auch studiengangspezifische Verteilungen, die eine getrennte Analyse von Lehramts- und Diplom-/ bzw. Magisterabschlüssen erforderlich machen.

Seit spätestens Anfang der 1970er Jahre herrscht eine stabile Notenhierarchie an deutschen Hochschulen. Die Rangfolge beginnt mit Biologie als Studiengang mit den besten Durchschnittsnoten, dicht gefolgt von Psychologie. In den ersten juristischen Staatsexamen werden im Mittel die schlechtesten Noten vergeben, sie sind mehr als doppelt so hoch wie in Biologie. In den beiden wirtschaftswissenschaftlichen Studiengängen BWL und VWL sind die schlechtesten Durchschnittsnoten innerhalb der Diplomstudiengänge zu finden. Jura, BWL und VWL decken über den gesamten Zeitverlauf betrachtet einen komplett anderen Bereich des Notenspektrums ab als Biologie, Psychologie, Mathematik und Chemie. Die Studiengänge mit sehr guten und sehr schlechten Notenniveaus weisen eine niedrige Streuung der Noten auf, für die Studiengänge im mittleren Bereich der Notenskala zeigt sich kein Muster.

Anhand der Anzahl an Jahren, in denen sich das Notenniveau zwischen den einzelnen Studiengängen signifikant unterscheidet, lassen sich aus der Rangfolge der 12 Studiengänge sieben Positionen bestimmen, die ein ähnliches Notenniveau aufweisen. Die Differenzen im Notenniveau zwischen diesen Positionen müssen jedoch stets im zeitlichen Kontext betrachtet werden - ihre zeitliche Stabilität variiert deutlich. Die Grenze zwischen gemeinsamem und signifikant unterschiedlichem Notenniveau liegt im Bereich 0.22 bis 0.24 Noten Abstand. Insgesamt lassen sich drei unterschiedliche Beziehungsmuster zwischen den einzelnen Studiengängen ausmachen: 1. Hohe Anzahl signifikant differenter Jahre und hohe zeitliche Stabilität dieser Differenzen, 2. Niedrige Anzahl signifikant differenter Jahre und niedrige oder keine Stabilität dieser Differenzen und 3. Niedrige bis mittelhohe Anzahl signifikant differenter Jahre und niedrige bis mittlere Stabilität dieser Differenzen. Auffällig ist, dass die beiden in der Stichprobe enthaltenen Magister- und die beiden Lehramtsstudiengänge jeweils ein ähnlich hohes Notenniveau aufweisen.

Aus der langfristigen Perspektive lassen sich die Notenentwicklungen auf Studiengangebene in zwei unterschiedliche Verlaufsformen einteilen: Entweder sie verbessern sich langfristig, begleitet durch zyklische Schwankungen oder sie verlaufen zyklisch auf einem relativ stabilen Notenniveau. Die Diplomstudiengänge weisen mit Ausnahme vom Studiengang Maschinenbau, für den aber auch nur Daten von zwei Hochschulen vorliegen, alle langfristige Verbesserungen auf, ebenso die beiden Lehr-

amtsstudiengänge. Die beiden Magisterstudiengänge und das erste juristische Staatsexamen hingegen bilden die Gruppe ohne langfristige Notenverbesserungen. In allen Studiengängen mit langfristiger Verbesserung setzt die erste Verbesserungsperiode in etwa zeitgleich zu Beginn/Mitte der 1960er Jahre ein und endet Anfang der 1970er Jahre, die zweite Verbesserungsphase beginnt jeweils im Laufe der 1980er Jahre, wobei sich das Ausmaß der Verbesserung und das Niveau auf dem sie sich vollzieht, studiengangspezifisch unterscheiden. Die Verbesserung der Noten wird durch Plateauphasen unterbrochen, die immer dann entstehen, wenn ein relativ schwacher Abwärtstrend auf eine Aufwärtsbewegung der zyklischen Zeitreihenkomponente trifft.

Die in allen Studiengängen zu beobachtenden Zyklen sind von ca. 20 Jahren Länge. In BWL und VWL, den Studiengängen mit dem schlechtesten Notenniveau der Studiengänge mit langfristiger Verbesserung nimmt die Stärke des Trends im Zeitverlauf zu. In Psychologie und Biologie ist die Verbesserung bereits nach der ersten Phase weitestgehend abgeschlossen, da die Noten kaum noch besser werden können, die Stärke des Trends nimmt entsprechend ab. Hier überdauert die Plateauphase die Aufwärtsphase der zyklischen Schwankung, seit Beginn der 1970er Jahre kann dort von grade compression gesprochen werden. In allen Studiengängen mit Verbesserung sinkt auch die Streuung der Noten im Zeitverlauf.

In den Studiengängen mit zyklischem Verlauf auf relativ stabilem Notenniveau verlaufen die Noten in unterschiedlichen Schwankungsbreiten. Die Auf- und Abwärtsbewegungen der Noten beginnen und enden leicht versetzt, die Zyklen dauern aber wie auch in den Studiengängen mit Verbesserung ca. 20 Jahre.

Erklärungsbedürftige Differenzen im Notenniveau und mehr oder weniger stabile Unterschiede in der langfristigen Entwicklung der Noten existieren nicht nur zwischen den Studiengängen sondern auch innerhalb dieser zwischen einzelnen Hochschulen. Auch dort verlaufen die langfristigen Verbesserungen nicht unbedingt auf dem gleichen Niveau und in gleichem Ausmaß, setzen die Verbesserungen nicht überall gleichzeitig ein und verlaufen die Zyklen teils in unterschiedlich langer Dauer und in unterschiedlichen Notenhöhen.

FH4a (an einzelnen Hochschulen existieren signifikante, im Zeitverlauf stabile Abweichungen vom durchschnittlichen Notenniveau im jeweiligen Studiengang) kann damit bestätigt werden. Die Streuung der Noten fällt im selben Studiengang ebenfalls teils unterschiedlich stark aus. Gemeinsamkeiten finden sich in den Hochschulen vor allem hinsichtlich der langfristigen Entwicklung: In der Regel sind auch an den einzelnen Hochschulen zyklische Bewegungen von 10-20 Jahren Länge zu finden, die entweder einen langfristigen Abwärtstrend begleiten oder sich relativ gleichmäßig um ein konstantes Notenniveau herum bewegen. Die Abwärtstrends weisen in den meisten Studiengängen mit langfris-

tiger Verbesserung einen einigermaßen parallelen Zeitraum auf, in dem der Trend am stärksten ist - auch wenn vereinzelt Abweichungen beim Beginn oder Ende dieser Zeiträume zu beobachten sind.

8.2 Einflüsse auf die Notengebung an deutschen Hochschulen

8.2.1 Leistungskonforme Prüfungsbedingungen

Lehrqualität

Unterschiedliche Notenniveaus können am einfachsten durch unterschiedliche Leistungen erklärt werden. Unterschiedliche Leistungen wiederum beruhen entweder auf unterschiedlichen Leistungsvoraussetzungen der Studierenden oder auf unterschiedlicher Lehrqualität.

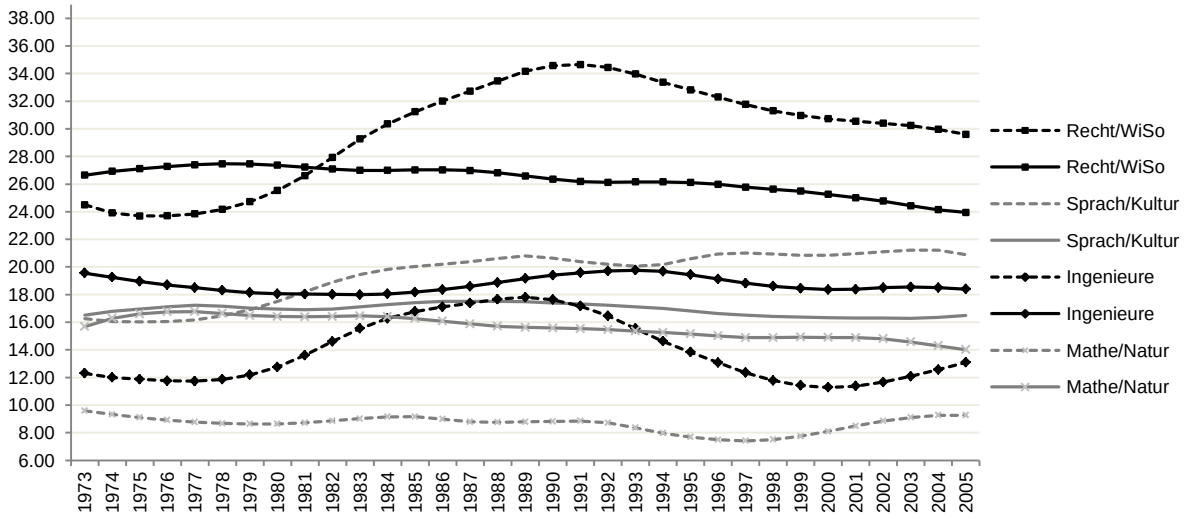
Als Indikatoren für die Lehrqualität auf der Makroebene können die personelle und finanzielle Ausstattung einer Hochschule oder eines Studiengangs betrachtet werden: Je besser die personelle Ausstattung, umso intensiver können die Studierenden betreut werden, je besser die finanzielle Ausstattung, umso besser können die materiellen Lehrbedingungen (z.B. die technische Ausstattung) gestaltet werden und umso eher besteht die Möglichkeit, Wissenschaftler*innen mit hoher Reputation als Lehrende zu gewinnen bzw. zu halten. Sowohl die personellen als auch die finanziellen Mittel müssen jedoch stets in Abhängigkeit der Anzahl der Studierenden betrachtet werden, auf die sie angewendet werden müssen. Mindestens die Finanzmittel können auch dann noch nicht ohne weiteres zwischen beliebigen Studiengängen verglichen werden - in den reinen Wissenschaften, etwa in der Philosophie, wo hinsichtlich der Ausstattung vor allem Bedarf an Literatur besteht, werden zur Durchführung der Lehre etwa weniger Mittel benötigt als in den angewandten, etwa in der Chemie, in der die Lehre auf kostspielige Laborausstattungen angewiesen ist.

Als Indikator für die jeweiligen Lehrbedingungen aussagekräftiger wären Vergleiche der finanziellen Mittel verschiedener Hochschulen innerhalb eines Studiengangs. Leider steht jedoch die Hochschulfinanzstatistik, die seit 1992 Informationen zur Höhe der finanziellen Mittel auf Hochschulebene enthält, der Forschung zum jetzigen Stand nicht im nötigen Format zur Verfügung, um entsprechende Analysen durchführen zu können.

Längsschnittdaten zu den Betreuungsrelationen an deutschen Hochschulen liegen nur nach Fächergruppen aufgegliedert vor, für den Zeitraum von 1973 bis 2005 im Zeitreihenformat (Lundgreen et al. 2009). Die Abstufung der Fächergruppen nach Betreuungsrelation entspricht durchaus den Abstufungen der Studiengänge im Notenniveau: Das beste Betreuungsverhältnis findet sich in der Gruppe Mathematik/Naturwissenschaften, das schlechteste in der Gruppe Rechts- Wirtschafts- und Sozialwissenschaften. Die Ingenieurs- sowie die Sprach- und Kulturwissenschaften liegen dazwischen, mit leicht günstigerem Verhältnis bei den Ingenieur*innen.

Werden aus den Noten der einzelnen Fächer ungewichtete Durchschnitte auf Fächergruppenebene berechnet, ergeben sich die schlechtesten Fächergruppennoten entsprechend den schlechtesten Betreuungsrelationen in den Rechts-/Wirtschafts- und Sozialwissenschaften, die besten Noten entsprechend den besten Betreuungsrelationen in der Gruppe der Mathematik und Naturwissenschaften. Die Noten in den Ingenieurswissenschaften sind etwas schlechter als in der Gruppe der Sprach- und Kulturwissenschaften, obwohl das Betreuungsverhältnis bei Letzteren schlechter ist (Abb.124). Für beide Gruppen sind die Noten auf Fächergruppenebene allerdings auch nicht unbedingt aussagekräftig: In den Ingenieurswissenschaften sind im sample nur die (selbst dort nicht verlässlichen) Noten der Maschinenbaustudierenden enthalten, die Betreuungsrelation hingegen umfasst noch fünf weitere große ingenieurswissenschaftliche Studiengänge. Auch für die Sprach- und Kulturwissenschaften liegen in der Stichprobe nur Noten für zwei (Germanistik und Psychologie) der 11 großen Studiengänge, für die die Betreuungsrelation berechnet wurde, vor. In der Gruppe Mathematik/Naturwissenschaften sind mit Mathematik, Biologie und Chemie immerhin drei von sechs großen Studiengängen abgedeckt, in den Rechts-/Wirtschafts- und Sozialwissenschaften sogar vier von fünf.

Abbildung 124: Fächergruppen-Noten*10 (durchgehend) und Betreuungsrelationen (gestrichelt) - Zeitverlauf (LOWESS 0.3)



Quelle: Lundgreen et al. 2009,eigene Darstellung

Tabelle 59: OLS-Regression der Fächergruppennote auf die Betreuungsrelation

AV: Note der Fächergruppe	Koeffizient	Standardfehler	t-Statistik	P>|t|
Betreuungsrelation	0.069	0.004	17.48	0.000
Konstante	0.837	0.077	10.81	0.000
				n=132; r^2=0.68

Eine Regression der fächergruppenspezifischen Note auf die Betreuungsrelation zeigt einen hochsignifikanten positiven Zusammenhang (Tab.59). Die Lehrqualität und damit die Betreuungsrelation kann zum Zeitpunkt der Prüfung aber keinen Einfluss mehr auf die Leistungsentwicklung haben , sondern müsste sich im Studienverlauf, also im Zeitraum zwischen Studienbeginn und Abschlussprüfung, auswirken. Dies muss in den Analysen berücksichtigt werden, was durch eine Zeitverschiebung der Notenwerte um bis zu eine Studiendauer (ca. 5 Jahre) in die Vergangenheit (‚time lag') möglich

ist. Da aber auch bei einer solchen Verschiebung die Abstufung der Fächergruppen sowohl bezüglich der Noten als auch bezüglich der Relationswerte gleich bleibt, ändern sich die Regressionsergebnisse kaum (Die Koeffizienten sinken leicht, liegen zwischen 0.066 (Lead5) und 0.069 (Lead2), in allen Fällen ist p=0.000).

Ein Zusammenhang zwischen Betreuungsrelation und Höhe des Notenniveaus scheint mit den fächergruppenspezifischen Daten also durchaus konform. Sollte tatsächlich ein genereller Zusammenhang zwischen den Betreuungsverhältnissen und der Höhe der Noten bestehen, müsste er sich als langfristig wirkender Einfluss allerdings auch im Zeitverlauf zeigen: Sinkende Relationen müssten innerhalb der Studiengänge mit sinkenden Noten einhergehen und umgekehrt.

Dies zu überprüfen, ist anhand der vorliegenden Daten aufgrund des hohen Aggregatniveaus allerdings nur ansatzweise möglich. Während die Differenz in Notenhöhe und Betreuungsrelation im aggregierten Querschnitt zwischen den Fächergruppen größer ist als innerhalb der Fächergruppen, sind die Differenzen in der Entwicklung der Studiengänge, zumindest was die Noten betrifft, in der Regel zu groß, als dass mithilfe derart grob gefasster Daten ein möglicher Einfluss im Zeitverlauf überprüft werden könnte: Mit Psychologie und Germanistik einerseits sowie Jura und den wirtschaftswissenschaftlichen Studiengängen andererseits sind in zwei der vier Fächergruppen sowohl Studiengänge mit Notenverbesserung als auch ohne enthalten. Da die Daten in Maschinenbau zudem nur begrenzt aussagekräftig sind, bleibt nur die Gruppe Mathematik/Naturwissenschaften.

Unter den Annahmen, dass a) sich auch in den im sample nicht enthaltenen Studiengängen, die in der Betreuungsrelation enthalten sind (Physik/Astronomie, Pharmazie und Geographie), die Noten im Zeitverlauf verbessern und b) die Betreuungsrelation sich in allen Studiengängen der Fächergruppe einigermaßen parallel entwickelt, müsste sich dort bei einem positiven Zusammenhang zwischen Betreuungsverhältnis und Notenhöhe auch ein entsprechender Effekt im Zeitverlauf beobachten lassen. Unabhängig vom Zeittrend müssten die Noten immer dann am besten sein, wenn die Betreuungsrelation einige Semester zuvor am besten war und umgekehrt. Um den Zeittrend auszuschalten wird das Jahr, für das die Note jeweils berechnet wurde als unabhängige Variable mit in das Regressionsmodell aufgenommen. Da die Zeitreihen der Noten autokorreliert sind, werden für die Berechnung Prais-Winsten-Regressionsmodelle (P-W-Regression) verwendet (Prais/Winsten 1954; zum Vorteil von GLS (Generalized Least Squares) - gegenüber OLS-Regressionen bei Autokorrelation siehe außerdem Kadiyala 1968 und Wooldridge 2009).

Es zeigt sich innerhalb der Fächergruppe kein Zusammenhang zwischen der Betreuungsrelation und dem Notenniveau. Einzig die Zeitvariable weist in allen Modellen (Lead0 bis Lead5) einen hochsignifikanten Einfluss auf, das Modell mit Lead5 weist die höchste Güte auf (Tab.60). Da die Interpretation dieses Ergebnisses als Befund gegen die These der Auswirkung der Betreuungsrelation auf die Noten von zwei spekulativen Annahmen abhängig ist, ist sie jedoch nur in begrenztem Maße aussagekräftig.

Tabelle 60: P-W-Regression der Fächergruppennote auf die Betreuungsrelation in Mathematik/Naturwissenschaften

AV: Note Mathe/Natur (Lead5)	Koeffizient	Standardfehler	t-Statistik	P>|t|
Betreuungsrelation	-0.001	0.008	-0.10	0.919
Jahr	-0.010	0.001	-14.76	0.000
Konstante	21.25	1.360	15.62	0.000
n=33; r^2adj=0.92; Durbin-Watson OLS:1.593; Durbin-Watson P-W: 1.907				

Zusammensetzung der Studierenden

Die Zusammensetzung der Studierenden kann als kompositioneller Faktor auf die Höhe von Durchschnittsnoten einwirken: Höhere Anteile von Studierenden mit bestimmten Merkmalen, deren Gruppe im Mittel bessere Leistungen erzielt, können das durchschnittliche Notenniveau in bestimmten Studiengängen oder an bestimmten Hochschulen anheben, ohne dass dafür die Benotungsstandards sinken müssen. (Selbst-)Selektionsprozesse bestimmter Gruppen in bestimmte Studiengänge oder an bestimmte Hochschulen könnten zu Differenzen im Querschnitt, wie auch zu studiengang- und/oder hochschulspezifische Entwicklungen der Noten führen. Als kompositionelle Einflüsse sind entsprechend dem besprochenen Forschungsstand denkbar: Das Geschlecht, das Alter, die soziale Herkunft, die Ethnizität, die Studienerfahrung sowie der Status als Transferstudierende*r, Stipendiat*in und/oder Teilzeitstudierende*r.

Ein Einfluss des Anteils Teilzeitstudierende (schlechtere Noten bei höherem Anteil aufgrund vermehrter außeruniversitärer Verpflichtungen und daher geringeren Möglichkeiten des Engagements für das Studium - vgl. Behr/Theune (2016), die zeigen, dass mit höherer außeruniversitärer Arbeitsbelastung eine längere Studiendauer einhergeht) dürfte im deutschen Hochschulsystem kaum existieren. Laut der 20. Sozialerhebung des Deutschen Studentenwerks war im Sommersemester 2012 nur ein Prozent der Studierenden im Erststudium in Teilzeit eingeschrieben (Middendorfff et al. 2013).

Bei Bargel/Bargel (2014) liegt der Anteil Teilzeitstudierende (nicht beschränkt auf das Erststudium) im Wintersemester 2010/11 insgesamt bei 5.6%. Die Differenz ist durch die unterschiedliche Operationalisierung zu erklären - der Wert von Bargel und Bargel liegt nur noch leicht über dem Wert der Sozialerhebung von vier Prozent, wenn dort die berufsbegleitenden Studiengänge miteinbezogen werden und nicht nur das Erststudium betrachtet wird. Die Verteilung über die Fächer(gruppen) zeigt bei Bargel und Bargel für die in der Stichprobe vertretenen Fächergruppen einen relativen Maximalwert von 14.8% in den Wirtschaftswissenschaften und einen Minimalwert von 2.2% in den Ingenieurswissenschaften. Davon ausgegangen, dass Teilzeitstudierende aufgrund ihrer höheren Belastung neben dem Studium im Mittel mit einer ganzen Note schlechter abschließen als Vollzeitstudierende, würde diese Differenz von 12.6 Prozentpunkten zu einer Differenz von 0.126 Noten führen.

Damit ließen sich immerhin 26.3% des tatsächlichen Unterschieds im Notenniveau im Jahr 2010 erklären - vorausgesetzt, das Notenniveau in den Studiengängen, die bei Bargel und Bargel in den Fächergruppen enthalten sind, ist genauso hoch wie in den Studiengängen der jeweiligen Fächergruppe, die im verwendeten sample vorliegen. Da allerdings auch die Wahrscheinlichkeit des Studienab-

bruchs bei Teilzeitstudierenden deutlich höher liegt als bei Vollzeitstudierenden - der Anteil, derjenigen, die einen Abbruch erwägen liegt etwa doppelt so hoch (ebd.) - und damit die Anzahl Teilzeitstudierende im Erststudium deutlich überschätzt wird, ist davon auszugehen, dass der potentielle Erklärungsanteil unterschiedlich hoher Teilzeitstudierendenanteile zum Prüfungszeitpunkt deutlich geringer liegt.

Grundsätzlich ist allerdings eine weitgehende Übereinstimmung zwischen den Rangfolgen des Notenniveaus auf Fächergruppenebene und des Anteils Teilzeitstudierende in den einzelnen Paarvergleichen festzustellen. Nur Maschinenbau fällt mit einem im Vergleich zu den anderen Studiengängen/Fächergruppen niedrigem Anteil Teilzeitstudierende bei gleichzeitig relativ schlechtem Notenniveau aus der Reihe.

Tabelle 61: Anteil der Teilzeitstudierenden und durchschnittliche Abschlussnoten

Fächergruppe[a]	Prozentualer Anteil Teilzeitstudierende[b] (Rang)	Durchschnittliche Abschlussnote[c] (Rang)	Übereinstimmende Rangfolge
Wirtschaft[1]	14.8 (1)	2.28 (5)	4/5
Recht	9.8 (2)	3.30 (6)	4/5
Sprache/Kultur[2]	4.0 (3)	1.77 (3)	4/5
Sozialw./Politik[3]	4.0 (3)	1.73 (2)	4/5
Mathematik/Naturw.[4]	3.6 (5)	1.40 (1)	4/5
Ingenieursw.[5]	2.2 (6)	1.80 (4)	2/5

[a] In der Stichprobe enthalten diese Fächergruppen: [1] VWL, BWL, [2] Germanistik, [3] Soziologie, [4] Mathematik, Chemie, Biologie, [5] Maschinenbau
[b] Quelle: Bargel/Bargel (2014): Studieren in Teilzeit und Teilzeitstudium - Definitionen, Daten, Erfahrungen, Positionen und Prognosen
[c] Stichprobendaten, Durchschnittswerte der in der Stichprobe enthaltenen Studiengänge dieser Fächergruppe für 2010 (Jura für 2007)

Im Zeitverlauf müsste der Anteil Teilzeitstudierende bei schlechteren Leistungen zunehmend abgenommen haben, um verbessernd auf die Durchschnittsnoten einwirken zu können. Leider wurden die offiziell als Teilzeitstudierende Immatrikulierten in der 20. Sozialerhebung erstmals als eigene Gruppe erfasst. Da ein offizielles Teilzeitstudium noch 1996 nur in Berlin möglich war (Wissenschaftsrat 1998a), ist der Zeitraum, in dem eine Veränderung des Anteils Teilzeitstudierende Auswirkungen auf die Notenhöhe gehabt haben könnte, jedoch ziemlich klein. Und gemessen am Zeitbudget („de facto Teilzeitstudium“ (Isserstedt et al. 2010:343), verstanden als Studium mit einem Studienaufwand von weniger als 25 Stunden pro Woche) ist der Anteil Studierende mit geringem zeitlichen Engagement an allen Studierendem im Erststudium zwischen 1988, dem erstmaligen Erfassungsjahr, und 2006 über alle Erhebungswellen hinweg gestiegen (ebd.). Ein notenverbessernder Effekt ist also auszuschließen, wird nicht davon ausgegangen, dass Teilzeitstudierende mit besseren Noten abschließen als Vollzeitstudierende.

In welchem Maße Unterschiede auf Hochschulebene im selben Studiengang durch eine Differenz im Anteil Teilzeitstudierende möglich sind, sollten sie ihr Studium tatsächlich schlechter abschließen und welche Entwicklung die Anzahl Teilzeitstudierende dort genommen hat, ist aufgrund fehlender Daten nicht zu sagen.

Vier Prozent der Studierenden, hier der Normalstudierenden (Vollzeit-Erststudium, ledig, nicht mehr im Elternhaus wohnhaft) gaben in der Sozialerhebung 2012 an, regelmäßig durch Stipendien gefördert zu werden. Diese vier Prozent wurden maßgeblich durch Begabtenförderungswerke (47%) und durch das Deutschlandstipendium (22%) gefördert (Middendorfff et al. 2013). Ein Vergleich der Förderung durch das Deutschlandstipendium sowie durch die Studienstiftung des Deutschen Volkes (SDV), dem größten der derzeit dreizehn aktiven Begabtenförderungswerke, nach Fächergruppen, zeigt folgende Unterschiede: Beim Deutschlandstipendium stellen 2014 die Ingenieurswissenschaften mit 29% anteilig die meisten Geförderten, es folgen die Rechts-, Wirtschafts- und Sozialwissenschaften mit 26%. Mathematik und Naturwissenschaften liegen bei 20% und zu 11% stammten die Stipendiat*innen aus der Gruppe der Sprach- und Kulturwissenschaften (Bundesministerium für Bildung und Forschung 2015). Die SDV hingegen förderte 2014 nur 11% in den Ingenieurs-, dafür 19% in den Sprach- und Kulturwissenschaften. In den Gruppen Mathematik/Naturwissenschaften und Rechts-, Wirtschafts- und Sozialwissenschaften lagen die Anteile mit 20% bzw. 23% in der Höhe gleich hoch bzw. etwas unter denen des Deutschlandstipendiums (Studienstiftung des deutschen Volkes 2015).
Werden die Anteile der beiden Förderungseinrichtungen ungewichtet gemittelt, sind, unter der Annahme, dass diese fächergruppenspezifische Verteilung den Verteilungen der übrigen Förderungseinrichtungen entspricht, unter den ca. 100 000 Stipendiat*innen (vier Prozent von ca. 2.5 Millionen Studierenden insgesamt[80]) 15 000 aus den Sprach- und Kulturwissenschaften, jeweils 20 000 aus der Gruppe Ingenieurswissenschaften und Mathematik/Naturwissenschaften und 24 500 aus den Rechts-, Wirtschafts- und Sozialwissenschaften. In Relation zu den fächergruppenspezifischen Studierendenzahlen gesetzt, ergeben sich daraus folgende Anteile Geförderte an allen Studierenden der jeweiligen Fächergruppe: Sprach- und Kulturwissenschaften: 3.1%, Ingenieurswissenschaften: 4.0%, Mathematik/Naturwissenschaften: 4.4%, Rechts-, Wirtschafts- und Sozialwissenschaften: 3.2%. Die Differenz zwischen dem höchsten Anteil Geförderte in der Gruppe Mathematik/ Naturwissenschaften und dem niedrigsten Anteil in der Gruppe Sprach- und Kulturwissenschaften beträgt damit 2012 1.3 Prozentpunkte.
Angenommen, die Stipendiat*innen in beiden Fächergruppen seien durch die Förderung von jeder Notwendigkeit, Erwerbsarbeit zu leisten, befreit, konzentrierten sich voll auf ihr Studium und beendeten ihr Studium alle mit einem ‚sehr gut' , während die übrigen Studierenden im Durchschnitt eine ganze Note schlechter abschnitten[81]: Der unterschiedlich hohe Anteil an Studierenden mit Förderung

[80] Die hier verwendeten Studierendenzahlen, insgesamt wie fächergruppenspezifisch sind der Publikation „Bildung und Kultur - Studierende an Hochschulen (Fachserie 11, Reihe 4.1)" des Statistischen Bundesamtes entnommen.
[81] Über alle Befragten aller Erhebungswellen, die bereits die Zwischenprüfung in ihrem Studiengang abgelegt haben, zeigt sich in den Daten des Konstanzer Studierendensurveys (s. Fußnote 83) ein um 0.49 Noten besseres Abschneiden von Stipendiat*innen (n=1 632) gegenüber nicht Geförderten (n=32 754) in der Zwischenprüfung. Dieser hochsignifikante Mittelwertunterschied (p=0.000) liegt bei der Hälfte der hier angenommenen Differenz

würde dann gerade einmal zu einem Unterschied von 0.013 Noten im Niveau der beiden Fächergruppen führen.

Auf Hochschulebene ist der Einfluss unterschiedlich hoher Stipendiat*innenanteile als ähnlich gering einzuschätzen: Die SDV förderte 2014 2.07% der Studierenden an der Universität Heidelberg, an keiner Universität lag der Anteil ähnlich hoch. Unter der Annahme, dass sich die 2.07% SDV-Geförderten an der Universität Heidelberg (628 Studierende) entsprechend der hochschulübergreifenden Verteilung der SDV-Geförderten nach Fächergruppen aufteilen, wären dies 69 in den Ingenieurs-, 119 in den Sprach- und Kulturwissenschaften, 126 in der Gruppe Mathematik/Naturwissenschaften und 144 in den Rechts-, Wirtschafts- und Sozialwissenschaften. In Relation zu den fächergruppenspezifischen Anteilen der Studierenden an der Universität Heidelberg[82] gesetzt, ergibt sich aus diesen Zahlen der höchste Gefördertenanteil mit 2.6% in der Gruppe Rechts-, Wirtschafts- und Sozialwissenschaften (Ingenieurswissenschaften werden in Heidelberg nicht angeboten). Die TU Braunschweig ist als einzige Hochschule aus der hier verwendeten Stichprobe nicht im Ranking der Hochschulen mit den meisten SDV-Geförderten aufgeführt. Doch selbst, wenn dort der Anteil bei 0% liegen sollte, würde dies eine maximale Notendifferenz von 0.026 zwischen den beiden Universitäten in den rechts-, wirtschafts- und sozialwissenschaftlichen Studiengängen bedeuten, wenn alle Stipendiat*innen mit einer ganzen Note besser abschließen als die übrigen Studierenden.

Auch der mögliche Effekt einer Zunahme des Anteils an Stipendiat*innen im Zeitverlauf auf die Notenhöhe lässt sich schnell abschätzen. Bei 4% Stipendiat*innen 2012 und angenommenen 0% 1960 ergäbe sich bei einem linearen Anstieg und einem im Durchschnitt um eine ganze Note besseren Abschluss eine jährliche (fachübergreifende) Verbesserung von 0.00077 Noten seit 1960. Ein genereller Einfluss des Stipendiat*innenanteils würde also im Zeitverlauf keinen merkbaren Effekt produzieren. Natürlich muss dabei in Betracht gezogen werden, dass sich die Anteile möglicherweise fachspezifisch entwickeln. Doch bei der geringen Gesamtzahl der heute Geförderten und dem langjährigen Wirken der wichtigsten Studienstiftungen, die nach dem zweiten Weltkrieg ihre Arbeit (wieder) aufnahmen, ist auch auf Studiengangebene nicht mit derart starken Zunahmen zu rechnen, dass ein nennenswerter Effekt auf die Notenhöhe auftritt.

Der Konstanzer Studierendensurvey[83] enthält fach-/ bzw. studiengangspezifische Daten zur Entwicklung der Stipendiat*innenanteile. Die Daten zeigen, dass der Anteil derjenigen, die ihr Studium teil-

---

in der Examensnote. Dafür muss in Betracht gezogen werden, dass sich der Notenvorteil von Stipendiat*innen fachspezifisch darstellt. Im Vordiplom weisen Chemiestudierende z.B. einen mehr als doppelt so hohen Mittelwertunterschied ($\bar{x}_1 - \bar{x}_2 = 0.72$; $n_1 = 696$; $n_2 = 31$) auf wie Betriebswirtschaftler*innen ($\bar{x}_1 - \bar{x}_2 = 0.29$; $n_1 = 2\,190$; $n_2 = 60$).

[82] Die fächergruppenspezifische Verteilung der Studierenden ist der Publikation „Studierendenstatistik Wintersemester 2014/2015“ der Ruprecht-Karls-Universität Heidelberg entnommen.

[83] Der Konstanzer Studierendensurvey erhebt regelmäßig die Studiensituation sowie studentische Erfahrungen und Orientierungen. Die hier genutzte Version umfasst die ersten 11 Wellen (st11w).

weise oder hauptsächlich durch ein Stipendium finanzieren in den meisten Studiengängen[84] seit Mitte oder Ende der 1990er über den Anteilen der vorherigen Befragungen liegt. Allerdings ist in keinem Studiengang ein zunehmender Anstieg zu beobachten. Teilweise weisen die Werte hohe Sprünge auf, was wohl darauf zurückzuführen ist, dass die Gruppe der Stipendiat*innen aufgrund ihres geringen Anteils an allen Studierenden in einer Befragung wie dem Konstanzer Survey kaum repräsentativ abzubilden ist. Um dennoch zumindest einen ungefähren Eindruck zu erhalten, inwiefern eine Veränderung des Stipendiat*innenanteils Auswirkungen auf die Notenhöhe haben könnte, bietet sich ein Vergleich der nach Dekaden aggregierten Werte an. Die Befragungen der Semester 1984/85, 1986/87 und 1989/90 wurden dazu als 1980er Jahre, die der Semester 1992/93, 1994/95 und 1997/98 als 1990er Jahre und die der Semester 2000/01, 2003/04, 2006/07 und 2009/10 als 2000er Jahre zusammengefasst.

Dabei zeigt sich in fünf Studiengängen in den 1990ern ein höherer Wert der Anteile Geförderte gegenüber den 1980ern, in den 2000ern allerdings wieder ein niedriger Wert als in den 1990ern, in einem Fall sogar niedriger als in den 1980ern. Ebenfalls in fünf Studiengängen liegt der Anteil in den 1990ern unter dem der 1980er, in den 2000ern dann höher als in den 1990ern und in vier davon dort auch höher als in den 1980ern. In einem Studiengang liegt der Anteil in den 1990ern höher als in den 1980ern und in den 2000ern höher als in den 1990ern, in einem genau umgekehrt. Die größte Steigerung ist in Psychologie zu verzeichnen, dort liegt der Anteil in den 2000ern bei 4.8% der Befragten, in den 1990ern lag er im Durchschnitt bei 2.2%. Diese Steigerung von 2.6 Prozentpunkten gegenüber der vorherigen Dekade würde bei einer durchschnittlichen Notendifferenz von einer ganzen Note zwischen Stipendiat*innen und Nicht-Stipendiat*innen (die in einem Studiengang mit grade compression wie Psychologie schon eine starke Annahme darstellt) lediglich eine Verbesserung von 0.026 Noten bedeuten. Die 4.8% Geförderten in Psychologie stellen gleichzeitig den Spitzenwert der Studiengänge (in allen drei Dekaden) dar - und selbst bei einer linearen Entwicklung des Stipendiat*innenanteils von 0% in 1965 bis 4.8% in 2010 wäre gerade mal eine jährliche Verbesserung von 0.00107 Noten zu verzeichnen. Bei einer tatsächlichen jährlichen Verbesserung von durchschnittlich 0.022 Noten in Psychologie wären damit gerade einmal 4.86% erklärt.

Auf Hochschulebene liegen zwar keine Daten zur zeitlichen Entwicklung vor, es ist jedoch aufgrund der zuvor beschriebenen Querschnittsdaten davon auszugehen, dass auch die regionalen Verteilungen der Stipendiat*innen innerhalb der Studiengänge nicht in einem Maße angestiegen bzw. gesunken sind, dass sie hochschulspezifische Entwicklungen erklären können.

[84] Erfasst über das 1. Studienfach. Unterschiedliche Abschlussarten werden erst seit der zweiten Erhebungswelle 1984/1985 abgefragt, weshalb im Folgenden immer erst ab diesem Zeitpunkt Vergleiche mit den Daten des Studierendensurveys erfolgen. Die Lehramtsstudiengänge umfassen in den im Folgenden verwendeten Vergleichsdaten Werte über die Ausbildungsgänge für alle Schulformen, nicht nur für das Gymnasium, wie in der Stichprobe.

So modellhaft diese Ausführungen auch sein mögen, verdeutlichen sie, dass die Entwicklung des Anteils an Stipendiat*innen, wie auch der an Teilzeitstudierenden, trotz nicht vorhandener Daten im Zeitreihenformat, die eine detaillierte Prüfung zuließen, nicht als nennenswerte Größe zur Erklärung jahrzehntelanger Verbesserungsprozesse gefasst werden kann. Um es noch einmal deutlich zu machen: Die ausgeführten Zahlenspiele sind kein Beleg dafür, dass der beschriebene kompositionelle Effekt nicht existiert und die gefundenen Muster der Notengebung nicht eventuell begünstigt. Insbesondere kurzfristige Einflüsse, die - womöglich zeitverzögert - mit punktuellen Erweiterungen der Begabtenförderung auftreten, sind denkbar, können aufgrund der mangelnden Datenlage allerdings nicht überprüft werden. Die Darstellung verdeutlicht lediglich, dass der Effekt alleine, sollte er existieren, einfach zu gering wäre, um maßgeblich zu den nachgewiesenen Differenzen und *langfristigen* Entwicklungen auf Studiengang- und Hochschulebene beigetragen zu haben.

Die Annahme, dass mit mehr Studienerfahrung auch bessere Noten einhergehen, da die Studierenden mit zunehmenden Fachsemestern zunehmende Prüfungsübung erhalten (Jewell/McPherson 2012) ist in Deutschland abschlussspezifisch zu bewerten. Vor allem für die neuen Abschlüsse Bachelor und Master, in denen vom ersten Semester an regelmäßig Prüfungsleistungen abgenommen werden, klingt sie plausibel. Auch in Studiengängen, in denen eine intensive systematische Prüfungsvorbereitung die Regel ist, etwa das wiederholte Schreiben von Probeklausuren, wie in Jura, führt zunehmende Studienerfahrung möglicherweise zu einem souveränen Umgang mit der Prüfungssituation und einer starken Internalisierung standardisierter Prüfungsschemata (Towfigh et al 2014).
Für die klassischen Diplom- und Magisterabschlüsse sowie für das Staatsexamen im Lehramt jedoch, in denen im Regelfall mit der Zwischen- und der Hauptprüfung zwei konzentrierte Prüfungsphasen zu absolvieren sind, hat das Argument der zunehmenden Prüfungseinübung keine Substanz. Hier wäre, ließe sich zeigen, dass höhere Fachsemester bessere Noten erzielen, eher von einem generellen Alterseffekt auszugehen. Die empirischen Befunde zum Zusammenhang zwischen Studiendauer und Studienerfolg, die sich finden lassen, weisen allerdings in die entgegengesetzte Richtung: Mit zunehmender Studiendauer geht eher ein verringerter Studienerfolg einher, sowohl in Bezug auf die Abschlussquote (Brinkmann 1967: ab einem bestimmten Schwellenwert, der über der Regelstudienzeit liegt) als auch auf die Notenhöhe (ebd.; Apenburg et al. 1976; Mosler/Savine 2004 für Vordiplomnoten; Ottwaska 1971; Wittenberg 2005). Der Zusammenhang zwischen zunehmender Studiendauer und schlechter werdenden Noten zeigt sich, zumindest in Informatik an der Leibniz Universität Hannover, auch im modularisierten Bachelorabschluss (v. Holdt et al. 2006) sowie im Master of Business Administration (an der Universität Potsdam (Madani et al. 2013)), was gegen die Hypothese der zunehmenden Prüfungseinübung spricht.

Es ist also umgekehrt eher davon auszugehen, dass ein schneller Abschluss des Studiums auch mit einem guten Notendurchschnitt einhergeht, wie es das aktuelle Forschungsverständnis, nach dem beide Variablen den Studienerfolg darstellen und daher auch gemeinsame Ursachen haben sollten, nahelegt. Ein verbessernder Einfluss einer verkürzten Studiendauer auf die Notenhöhe ist in dieser Kausalrichtung auch theoretisch nicht begründbar - hier ließe sich eher umgekehrt argumentieren, dass Studierenden, die bessere Noten erzielen, das Studium einfacher fällt, weshalb sie es auch schneller beenden.

Auch ein Alterseffekt kann theoretisch in zwei Richtungen wirken. Es lässt sich häufig die Vermutung finden, ältere Studierende würden die besseren Leistungen erzielen, da sie mehr Lebenserfahrung mitbringen, die ihnen hilft, die Lehr- und Lernstrukturen besser zu erfassen und sie das Studium aufgrund ihrer höheren Reife ernster nehmen als ihre jüngeren Kommiliton*innen (vgl. Erdel 2010; Jirjahn 2007; Kwon et al. 1997; Prather et al. 1979). In der US-amerikanischen Literatur gibt es nur geringe empirische Evidenz für diese These (siehe Tab.4). Für Deutschland weisen die wenigen empirischen Studien mit nur einer Ausnahme in die entgegengesetzte Richtung: Lediglich Jirjahn (2007) findet bessere Noten im Vordiplom bei älteren Studierenden der Wirtschaftswissenschaften an drei deutschen Hochschulen.

Grözinger (2015) stellt anhand der amtlichen Hochschulprüfungsstatistik fest, dass in den Wirtschaftswissenschaften abschlussübergreifend ein hochsignifikanter positiver Zusammenhang zwischen dem Alter und der Notenhöhe besteht, ältere Studierende also schlechtere Noten erzielen als jüngere. Mosler und Savine (2004) können zeigen, dass jüngere Studierende in VWL und BWL an der Universität zu Köln im Vordiplom besser abschneiden. Dies passt zu den Ergebnissen von Erdel (2010) nach denen jüngere Studierende die Grundlagenphase des Bachelors am Fachbereich Wirtschaftswissenschaften der Universität Erlangen-Nürnberg mit besseren Noten abschließen und zum älteren Befund von Hampe (1977; 1978), dass bei Fachstudienbeginn ältere (die Studie umfasst nur männliche) Studierende in Jura, VWL, Medizin und Zahnmedizin an der Universität Marburg am Ende des Studiums schlechtere Abschlussnoten erzielen.

Die Daten des Konstanzer Studierendensurvey deuten darauf hin, dass auch bezüglich des Alters fach- bzw. abschlussspezifische Zusammenhänge bestehen. Über alle Befragten (n=34 291) aller Abschlüsse und Studiengänge hinweg, die bereits eine Zwischenprüfung abgelegt haben, zeigt sich kein Zusammenhang zwischen den darin erzielten Noten und dem Alter der Befragten (r=0.004; p=0.419). Wird die Korrelation zwischen den beiden Variablen allerdings abschluss- und fachspezifisch berechnet, ergibt sich ein mittlerer positiver Zusammenhang (höheres Alter geht mit schlechterer Note einher) in den Diplomstudiengängen/-fächern (1. Studienfach) Psychologie (r=0.277; p=0.000; n=794), Maschinenbau/Produktions- u. Verfahrenstechnik (r=0.226; p=0.000; n=2 459) und Mathematik/Statistik (r=0.205; p=0.000; n=536). Im Diplom Biologie (r=0.167; p=0.000; n=728), (Lebensmittel-/

Bio-) Chemie (r=0.160; p=0.000; n=724), VWL (r=0.124; p=0.008; n=456), BWL (r=0.079; p=0.000; n=2 241) und im Staatsexamen Jura (r=0.083; p=0.028; n=699) ist dieser Zusammenhang nur in geringem Maße vorhanden. Die fachspezifische Abstufung der Effektstärke entspricht in etwa der, die Grözinger (2017) anhand der Hochschulprüfungsstatistik ermittelt. In den beiden Lehramts-, wie auch in den beiden Magisterstudiengängen besteht kein signifikanter Zusammenhang. Einschränkend ist darauf hinzuweisen, dass der Datensatz nur das Alter zum Befragungszeitpunkt, nicht zum Zeitpunkt der Zwischenprüfung oder zum Studienbeginn enthält und die Zwischenprüfungsnoten als Ergebnis einer vorgezogenen Selektionsstufe, die eine andere Funktion erfüllt als das Examen, nicht mit Abschlussnoten gleichzusetzen sind.

Bezüglich des ersten Punkts ergibt sich ein komplett neues Bild, wenn aus dem Befragungsalter und der Anzahl Fachsemester das Alter zu Studienbeginn errechnet wird: Lediglich in Psychologie (r=0.250; p=0.000; n=787), Biologie (r=0.124; p=0.001; n=722) und Maschinenbau/Produktions- u. Verfahrenstechnik (r=0.071; p=0.000; n=2 425) besteht ein signifikanter Zusammenhang zwischen dem Alter zu Studienbeginn und der Note in der Zwischenprüfung, nur in Psychologie ist er in etwa so stark wie zwischen dem Befragungsalter und der Note, in den beiden anderen Fächern deutlich niedriger. Der positive Zusammenhang zwischen Befragungsalter und Notenhöhe dürfte demnach eher durch schlechtere Leistungen bei längerer Studiendauer zustande kommen. Hierzu passt auch der notenverbessernde Effekt einer Berufsausbildung bei Grözinger (2017).

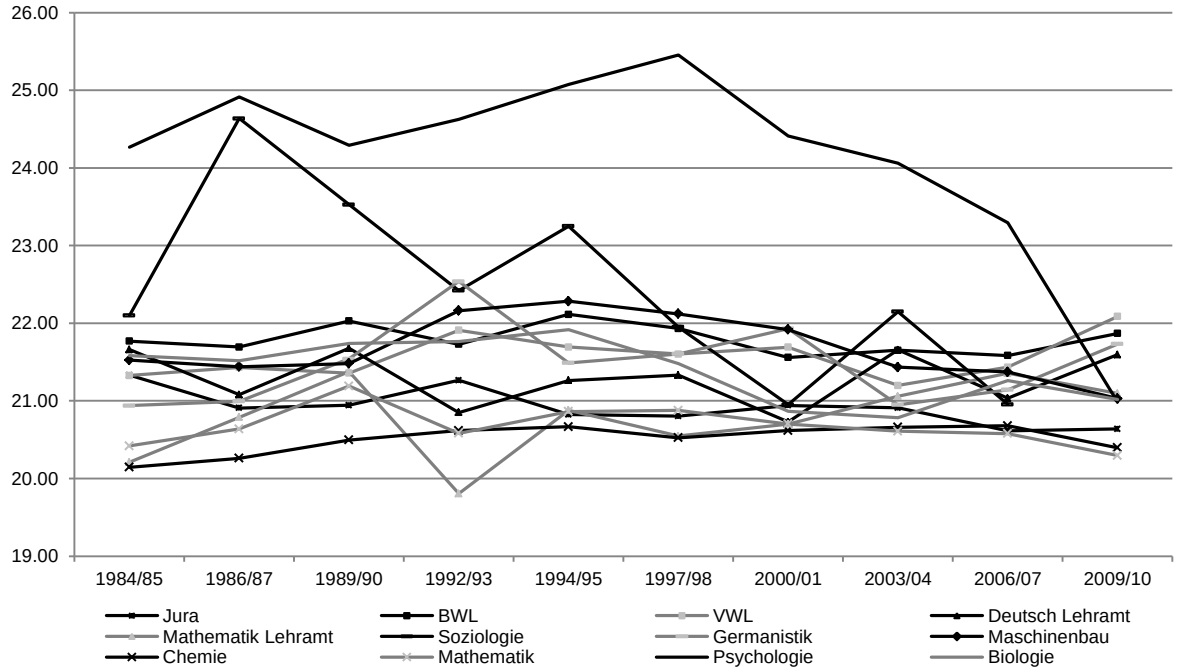
Der zweite Punkt betrifft auch die anderen vorgestellten empirischen Befunde mit Ausnahme von Hampe (1977; 1978). Grundsätzlich weisen die verfügbaren Daten jedoch darauf hin, dass sich der Zusammenhang zwischen Alter und erzielter Note, soweit im konkreten Studiengang überhaupt existent, eher so darstellt, dass jüngere Studierende bessere Noten erzielen als umgekehrt. Auch eine Regression der Zwischenprüfungsergebnisse auf das Alter zu Beginn des Studiums zeigt in den Daten des Studierendensurveys mit positivem, signifikantem Koeffizienten (b=+0.085; p=0.000) tendenziell in diese Richtung, weist aber keine Erklärungskraft auf (r^2=0.002).

Entsprechend ließe sich, wenn dann, erwarten, dass die Alterskomposition in Studiengängen mit einem niedrigeren Durchschnittsalter der Studierenden bessere Noten im Vergleich zu Studiengängen mit höherem Durchschnittsalter begünstigt. Über alle Erhebungswellen des Studierendensurveys gemittelt, weisen die Psycholog*innen mit 24.3 Jahren den höchsten Alterswert der in der Stichprobe enthaltenen Studiengänge bei Studienbeginn auf, die Studierenden aller chemischen Studiengänge waren bei Aufnahme ihres Studiums mit durchschnittlich 20.5 Jahren am jüngsten. Auch die Jurist*innen sind mit 21.0 Jahren vergleichsweise jung bei Studienbeginn. Dass die Psychologie als einer der beiden Studiengänge mit den langfristig besten Durchschnittsnoten in der Stichprobe die mit Abstand ältesten Studienanfänger*innen aufweist und in Jura relativ junge Erstsemester konstant die

schlechtesten Noten erhalten, spricht gegen einen relevanten Beitrag der Alterskomposition zur Erklärung von Unterschieden in der Notenhöhe.

Auch zur Erklärung der langfristigen Entwicklung der Noten kann die Alterskomposition nicht beitragen: Im Zeitverlauf ist der Anteil der älteren Studierenden fachübergreifend betrachtet tendenziell gestiegen (Lundgreen et al. 2008, Tab.2.32), die Noten hätten demnach im Zeitverlauf schlechter werden müssen, gäbe es einen generellen Alterseffekt. Fachspezifisch ist eine Abnahme des Alters zu Studienbeginn über die Erhebungswellen des Surveys tendenziell in Soziologie Magister und, schwächer ausgeprägt, in Jura bzw. ab Mitte/Ende der 1990er Jahre in Maschinenbau, Mathematik, Psychologie und Biologie zu beobachten (Abb.125). Damit sinkt das Alter sowohl in Studiengängen mit als auch in solchen ohne Notenverbesserung seit den 1980er Jahren - es sinkt aber nicht in den übrigen Studiengängen, die allesamt Verbesserungen im entsprechenden Zeitraum aufweisen.
Um mögliche fachspezifische Zusammenhänge dort zu überprüfen, wo ein tendenziell sinkendes Alter mit sich verbessernden Notendurchschnitten einhergeht, sind die jeweils 10 Messzeitpunkte nicht ausreichend. Auf Hochschulebene stehen keine Daten zum Vergleich der Altersdurchschnitte zur Verfügung. Da ein genereller Alterseffekt nicht festgestellt werden kann, ist allerdings auch auf Hochschulebene nicht damit zu rechnen, dass die Entwicklung des Durchschnittsalters zur Erklärung von Unterschieden im Notenniveau zwischen Hochschulen bzw. im Zeitverlauf geeignet ist.

Abbildung 125: Durchschnittliches Alter bei Studienbeginn



Quelle: Konstanzer Studierendensurvey 1983-2010, eigene Berechnungen. Der Datenpunkt 1989/90 wurde für Soziologie durch einen linear interpolierten Wert ersetzt, da der Originalwert von 29.71 Jahren als fallzahlenbedingter (n=14) Ausreißer betrachtet werden muss. Soziologie inkl. Sozialwissenschaften/Sozialkunde; Maschinenbau inkl. Produktions- und Verfahrenstechnik; Chemie inkl. Bio-/Lebensmittelchemie; Mathematik inkl. Statistik.

Die soziale Herkunft von Studierenden ist ein klassischer Faktor, an dem Chancenungleichheiten und damit auch unterschiedliche Erfolge im Studium assoziiert werden. Üblicherweise wird argumentiert,

dass Akademiker*innenkinder durch die Kenntnisse, die ihre Eltern vom Hochschulsystem besitzen, Vorteile in ihren Lernstrategien, in ihrer Kurswahl usw. erhalten (Jirjahn 2007). Zudem sinkt mit steigender Schichtzugehörigkeit einerseits die Notwendigkeit, Zeit für Erwerbsarbeit zur Finanzierung des Studiums aufbringen zu müssen, andererseits ist bei entsprechender finanzieller Absicherung im Elternhaus womöglich auch der Anreiz niedriger, das Studium zügig und mit einem guten Abschluss zu beenden (ebd.).

Empirisch lassen sich auch hier auf den ersten Blick gegensätzliche Befunde ausmachen, die sich allerdings fachspezifisch auflösen lassen. So können Ottwaska (1971), Hampe (1977; 1978 - auch für Medizin und Zahnmedizin) und Erdel (2010) in den von ihnen untersuchten wirtschaftswissenschaftlichen Studiengängen keinen begünstigenden Einfluss der akademischen Herkunft auf die Notenhöhe finden. Hampe (1977) stellt allerdings fest, dass Akademiker*innenkinder in Jura im Durchschnitt um 0.2 Noten besser sind als ihre Kommiliton*innen ohne akademischen Hintergrund.

Diese fachspezifische Differenzierung lässt sich auch im Konstanzer Studierendensurvey wiederfinden: Während über alle Befragten, die ihre Zwischenprüfung bereits abgeschlossen haben, Studierende mit mindestens einem akademischen Elternteil (x_1; n=15 853) im Mittel hochsignifikant (p=0.000) um 0.08 Noten besser in der Zwischenprüfung abschneiden, als solche ohne diesen Hintergrund (x_2; n=18 141), ist dieser Unterschied auf Studiengangebene nicht überall zu finden. Von den im sample enthaltenen Studiengängen/Fächern zeigt sich dieser signifikante Unterschied nur in Biologie ($\bar{x}_1$- $\bar{x}_2$=-0.20; p=0.000; n_1=346; n_2=378), (Bio-/Lebensmittel-)Chemie ($\bar{x}_1$- $\bar{x}_2$=-0.16; p=0.003; n_1=333; n_2=385), Mathematik/Statistik ($\bar{x}_1$- $\bar{x}_2$=-0.15; p=0.022; n_1=284; n_2=248), Psychologie ($\bar{x}_1$- $\bar{x}_2$=-0.14; p=0.002; n_1=392; n_2=394) und Jura ($\bar{x}_1$- $\bar{x}_2$=-0.09; p=0.038; n_1=418; n_2=274). In den beiden wirtschaftswissenschaftlichen Studiengängen erzielen Akademiker*innenkinder passend zu den vorgestellten Befunden keine (signifikant) besseren Noten, ebenso wie in den anderen Studiengängen des samples. In BWL sind ihre Vordiplomergebnisse mit einer Differenz von 0.02 Noten (p=0.396; n_1=964; n_2=1 254) sogar minimal schlechter. Der Effekt der Mittelwertdifferenz lässt sich folgendermaßen ablesen: Bei einer Gleichverteilung von Akademiker*innen- und Nichtakademiker*innen senkt (hebt) der Notenvorteil (der Notennachteil) einer der beiden Gruppen das Gesamtnotenniveau um die Hälfte der Mittelwertdifferenz gegenüber dem Niveau, das bei ebenfalls gleichen Noten von Prüflingen mit und ohne Akademiker*innenelternteil herrschen würde (Tab.62, Spalten 2 und 3).

Die Anteile der Akademiker*innenkinder an der Gesamtheit der Studierenden im Studiengang/Fach liegen zwischen 41.4% in BWL (alle Studierende) bzw. 40.7% in Maschinenbau/Produktions-und Verfahrenstechnik (nur Studierende, die die Zwischenprüfung abgelegt haben) und 55.9% bzw. 60.8% in Jura. Da das Ausmaß der Differenzen im Anteil alleine keine Unterschiede im Notenniveau produzieren kann, wenn die Akademiker*innenkinder nicht auch im Durchschnitt bessere Noten aufweisen,

ist immer die Kombination mit dem Notenvorteil relevant. Der Effekt der Verteilung der Bildungsherkunft in Kombination mit der festzustellenden Mittelwertdifferenz lässt sich folgendermaßen ablesen: Mit jedem Prozentpunkt, den der Anteil der besser (schlechter) bewerteten Gruppe steigt (und der der anderen Gruppe sinkt), verbessert (verschlechtert) sich das Gesamtnotenniveau um den Wert der Mittelwertdifferenz geteilt durch 100 (Spalten 4 und 5).
Sowohl in Chemie als auch in Mathematik erhalten die Akademiker*innenkinder zwar die besseren Noten, nur in Mathematik sind sie aber auch in der Mehrzahl, was sich im Vergleich zu Chemie positiv auf das Notenniveau auswirkt - denn dort sind die schlechter bewerteten Nicht-Akademiker*innenkinder in der Überzahl. In Relation der Studiengänge zueinander ergibt sich damit ein (in der Stärke in dieser Reihenfolge abnehmender) notensenkender Effekt in Mathematik Lehramt, Mathematik, Jura, Germanistik, Deutsch Lehramt und in Biologie, Chemie, Psychologie, Soziologie, VWL, Maschinenbau sowie in BWL ein (ebenfalls in dieser Reihenfolge abnehmender) notenhebender Effekt (Spalte 6).

Den größten absoluten Effekt auf die Notendifferenzen zwischen den einzelnen Studiengängen mit signifikanten Notenunterschieden erzielt die unterschiedliche Verteilung der beiden Gruppen in Kombination mit den jeweiligen Notenunterschieden dabei für die Differenz zwischen Mathematik und Biologie. Im Vergleich zur Differenz, die das Gesamtnotenniveau zwischen diesen beiden Studiengängen aufweisen würde, wenn die Anteile und Noten beider Gruppen gleich wären, ergibt sich eine *maximale*[85] Vergrößerung von 0.182 (0.078 Verbesserung in Mathematik + 0.104 Verschlechterung in Biologie) Noten. Diese maximale Vergrößerung der Differenz liegt deutlich über dem tatsächlichen Unterschied im aggregierten Notenniveau von 0.014 Noten.
Für das Paar Chemie-Biologie liegt der Anteil der maximalen Veränderung bei 31.7% der tatsächlichen Notendifferenz, für Psychologie-Biologie bei 14.9% und für Psychologie-Chemie bei 8.2%. Für die Paare Psychologie-Mathematik, Psychologie-Jura, Chemie-Mathematik, Chemie-Jura, Mathematik-Jura und Biologie-Jura tragen die Disparitäten hingegen nicht zur Erklärung der Differenzen bei - hier würde sich die Differenz im Notenniveau sogar noch vergrößern, wären die Anteile und Noten jeweils gleich verteilt.

Ohne eine multivariate Analyse des Effekts unter Kontrolle anderer möglicher Einflussfaktoren ist jedoch zu betonen, dass es sich bei den prozentualen Werten lediglich um die Maximalwerte der durch unterschiedliche notenverbessernde Einflüsse von Akademiker*innenkindern erklärbaren An-

[85] Hier gilt, dass die maximale Veränderung gegenüber der Annahme gleicher Anteile beider Gruppen mit gleichen Durchschnittsnoten nur erreicht wird, wenn sich die Veränderung, die sich bei gleich hohen Anteilen durch die Mittelwertdifferenz alleine ergibt, das Notenniveau in die gleiche Richtung verändert, in die auch die unterschiedlich hohen Anteile bei gegebener Differenz wirken.

teile an den Notendifferenzen handelt. Da auch Einflüsse existieren mögen, die die Notendifferenzen zwischen den Studiengängen erhöhen, kann also nicht gefolgert werden, dass eine Nivellierung der Notenvorteile von Akademiker*innenkindern etwa die Notendifferenz zwischen Chemie und Biologie um ein Drittel senken würde. Dennoch zeigen die Rechnungen, dass unterschiedliche Anteile Akademiker*innenkinder mit unterschiedlich großen Notenvorteilen zumindest in begrenztem Umfang zu Unterschieden im Notenniveau beitragen können.

Tabelle 62: Studiengangspezifische Mittelwertvergleiche (T-Test) der Zwischenprüfungsnoten zwischen Akademiker- und Nichtakademiker*innenkindern und Auswirkungen auf das Notenniveau

Studiengang	Aggregierte Mittelwertdifferenz[a]	Veränderung gegenüber gleicher Durchschnittsnote	Anteil Akademiker *innenkinder	Veränderung gegenüber gleichem Anteil	Maximale Veränderung
Biologie	0.197***	±0.099	47.7%	+0.005	+0.104
Mathematik Lehramt	0.188	±0.094	53.2%	-0.006	-0.100
Chemie	0.157**	±0.079	46.7%	+0.005	+0.084
Mathematik	0.145*	±0.073	53.5%	-0.005	-0.078
Psychologie	0.142**	±0.071	49.9%	+0.000	+0.071
Soziologie	0.119	±0.060	47.7%	+0.003	+0.063
Jura	0.095*	±0.048	60.8%	-0.010	-0.058
VWL	0.092	±0.046	42.4%	+0.007	+0.053
Germanistik Magister	0.077	±0.039	50.5%	-0.000	-0.039
Deutsch Lehramt	0.066	±0.033	54.9%	-0.003	-0.036
Maschinenbau	0.033	±0.017	40.7%	+0.003	+0.020
BWL	-0.019	±0.010	43.5%	-0.001	+0.011

*p≤0.05 **p≤0.01 ***p≤0.001

[a]Mittelwertdifferenz=Durchschnittliche Abschlussnote Nichtakademiker*innenkind - Akademiker*innenkind

Quelle: Konstanzer Studierendensurvey 1983-2010, eigene Berechnungen. Soziologie inkl. Sozialwissenschaften/Sozialkunde; Maschinenbau inkl. Produktions- und Verfahrenstechnik; Chemie inkl. Bio-/Lebensmittelchemie; Mathematik inkl. Statistik.

Im Längsschnitt kann nicht ausgeschlossen werden, dass ein steigender Anteil Akademiker*innenkinder sinkende Noten begünstigt, aber auch hier ist nur von einem sehr begrenzten Effekt auszugehen, sollte er existieren. Der Anteil Studierende mit mindestens einem akademischen Elternteil ist über die Wellen des Konstanzer Studierendensurveys - mit Ausnahme der beiden Lehramtsstudiengänge - in allen im sample enthaltenen Fächern/Studiengängen (in unterschiedlichem Maße) gestiegen. Der stärkste Anstieg ist in VWL zu verzeichnen[86], von 37.6% in 1985 auf 69.7% in 2010 (+32.1 Prozentpunkte).

Bei einem angenommenen Notenvorteil von einer ganzen Note gegenüber Nichtakademiker*innenkindern würde sich dieser Anstieg in einer Notenverbesserung von 0.321 Noten in 25 Jahren niederschlagen. Zumindest in der Zwischenprüfung, die im Survey abgefragt wird, beläuft sich der tatsächliche Notenvorteil der Akademiker*innenkinder in VWL (über alle Wellen berechnet) allerdings gerade einmal auf 0.09 Noten (der maximale Notenvorteil liegt bei 0.20 Noten in Biologie). Damit liegt die Verbesserung durch den gestiegenen Anteil gerade einmal bei 0.029 Noten in 25 Jahren (in Biologie als Studiengang mit der größten Mittelwertdifferenz führt der geringere Anstieg zu einer Verbesserung von 0.021 Noten im selben Zeitraum).

[86] Noch stärker ist der Anstieg in Soziologie (+48,6 Prozentpunkte). Aufgrund der dort sehr niedrigen Fallzahlen in den ersten Erhebungen (n=19 für die erste Welle), wird dieser Wert jedoch als nicht verlässlich eingestuft.

Wird der Bildungshintergrund durch die berufliche Stellung der Eltern (im Studierendensurvey nach Britt Hoffmann operationalisiert) ersetzt, ergibt sich über alle Befragten des Surveys, die bereits ihre Zwischenprüfung abgelegt haben, ein geringer negativer Zusammenhang zwischen beruflicher Position und Notenhöhe (n=33 775; Somers D=-0.047; p=0.000). Und auch unter Kontrolle des Zeittrends erzielen Studierende mit steigender beruflicher Stellung der Eltern zunehmend bessere Noten in der Zwischenprüfung (Tab.63, wieder nur Studierende der im sample enthaltenen Fächer/Studiengänge).

Tabelle 63: OLS-Regression der Zwischenprüfungsnote auf den Index der beruflichen Stellung

AV: Zwischenprüfungsnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Index berufliche Stellung	-0.022	0.004	-5.11	0.000
Jahr	-0.025	0.003	-9.23	0.000
Konstante	2.856	0.026	109.35	0.000
				n=9459; r^2adj=0.01

Studiengang-/fachspezifisch können signifikant bessere Zwischenprüfungsnoten bei höherer beruflicher Stellung für die Studiengänge der Stichprobe in Deutsch Lehramt (n=303 ; Somers D=-0.161; p=0.001), Mathematik (n=531; Somers D=-0.132; p=0.000), Germanistik Magister (n=310; Somers D=-0.113; p=0.016), Psychologie (n=781; Somers D=-0.110; p=0.000), Jura (n=690; Somers D=-0.107; p=0.001), Chemie (n=713; Somers D=-0.087; p=0.007) und Biologie (n=718; Somers D=-0.079 ;p=0.015) festgestellt werden. In Soziologie (n=132; Somers D=-0.073; p=0.313), Mathematik Lehramt (n=200; Somers D=-0.050; p=0.425), VWL (n=453; Somers D=-0.036; p=0.349) und Maschinenbau (n=2413; Somers D=-0.017; p=0.319) stimmt das Vorzeichen zwar mit den Erwartungen überein, jedoch ist der Zusammenhang nicht signifikant. In BWL findet sich sogar ein positives Vorzeichen (n=2215; Somers D=0.032; p=0.072). Ein Zusammenhang besteht damit in denselben Fächern/Studiengängen, in denen auch der akademische Bildungshintergrund notenverbessernd wirkt, zusätzlich aber auch in Germanistik Magister und Deutsch Lehramt.

Hinsichtlich der Erklärungskraft in Bezug auf die festgestellten Notendifferenzen lässt sich aus der unterschiedlichen Einflussstärke der beruflichen Stellung und dem unterschiedlichen Niveau abschätzen, inwiefern Notendifferenzen zwischen den Studiengängen durch die vorzufindenden Konstellationen begünstigt werden: Dazu lässt sich anhand der mittels linearer Regression geschätzten Veränderung der Zwischenprüfungsnote pro Indexstufe errechnen, um welchen Wert sich ein gegebenes Notenniveau bei Indexstufe 1 in den einzelnen Studiengängen/Fächern bei Erreichen des tatsächlich vorliegenden Indexwertes verbessert. Dabei wird deutlich, dass die Differenzen im Indexwert allein zu gering sind, um eine Rolle zu spielen. Der Einfluss auf das Notenniveau stuft sich exakt in der Reihenfolge der geschätzten Notenveränderung pro Indexstufe ab: In Mathematik übt die berufliche Stellung den größten notenverbessernden Effekt aus, in Jura ist der Effekt trotz des höchsten Indexwertes am geringsten.

Damit lässt sich parallel zum Bildungshintergrund theoretisch ein geringer Anteil der Notendifferenz zwischen Jura und den anderen Studiengängen/ Fächern mit notenverbesserndem Einfluss der beruflichen Stellung erklären, während die signifikante Notenverbesserung pro Indexstufe im Vergleich zu den Studiengängen/Fächern ohne signifikante Veränderung den Erklärungsbedarf vergrößert. Mit der Relation der Notendifferenz überein stimmt die Relation im Einfluss auf das Notenniveaus außerdem zwischen Deutsch Lehramt und a) Biologie, b) Chemie (beide nur in den Zwischenprüfungsnoten) sowie c) Mathematik (nur Abschlussnoten), zwischen Germanistik und a) Psychologie, b) Biologie, c) Chemie (alle nur Zwischenprüfungsnoten) sowie d) Mathematik (nur Abschlussnoten), zwischen Psychologie und a) Biologie (nur Zwischenprüfungsnoten) sowie b) Chemie, zwischen Biologie und a) Chemie (nur Abschlussnoten) sowie b) Mathematik (nur Zwischenprüfungsnoten) und schließlich zwischen Chemie und Mathematik (nur Abschlussnoten). Damit lässt sich auch für den beruflichen Hintergrund festhalten, dass eine Begünstigung von Notendifferenzen durch dieses Merkmal unterschiedlicher sozialer Herkunft plausibel ist, in seinem Umfang jedoch nur geringen Einfluss aufweist.

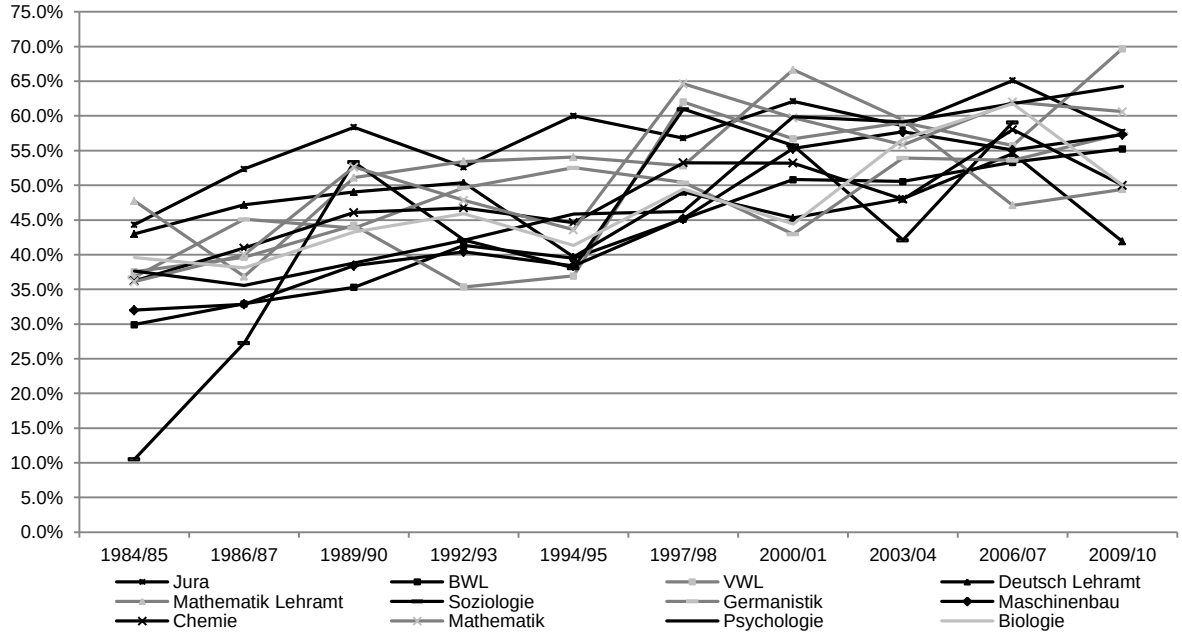
Tabelle 64: Mittlerer Indexwert (nur Zwischenprüfung bereits abgelegt) und Notenvorteil in der Zwischenprüfung

Studiengang/Fach	Mittlerer Indexwert	Notenveränderung pro Indexstufe	Einfluss auf das Notenniveau[a]
Jura	5.42	-0.036*	-0.159
Mathematik Lehramt	5.24	-0.032	--
Deutsch Lehramt	5.18	-0.064**	-0.268
Germanistik Magister	5.16	-0.051*	-0.212
Mathematik	5.12	-0.077***	-0.317
VWL	5.07	-0.024	--
Psychologie	5.05	-0.051**	-0.207
Biologie	4.95	-0.045**	-0.178
BWL	4.93	+0.009	--
Chemie	4.91	-0.043**	-0.168
Soziologie	4.87	-0.026	--
Maschinenbau	4.65	-0.006	--

[a] Im Vergleich zum Indexwert 1;*p≤0.05 **p≤0.01 ***p≤0.001;
Quelle: Konstanzer Studierendensurvey 1983-2010, eigene Berechnungen. Soziologie inkl. Sozialwissenschaften/Sozialkunde; Maschinenbau inkl. Produktions- und Verfahrenstechnik; Chemie inkl. Bio-/Lebensmittelchemie; Mathematik inkl. Statistik.

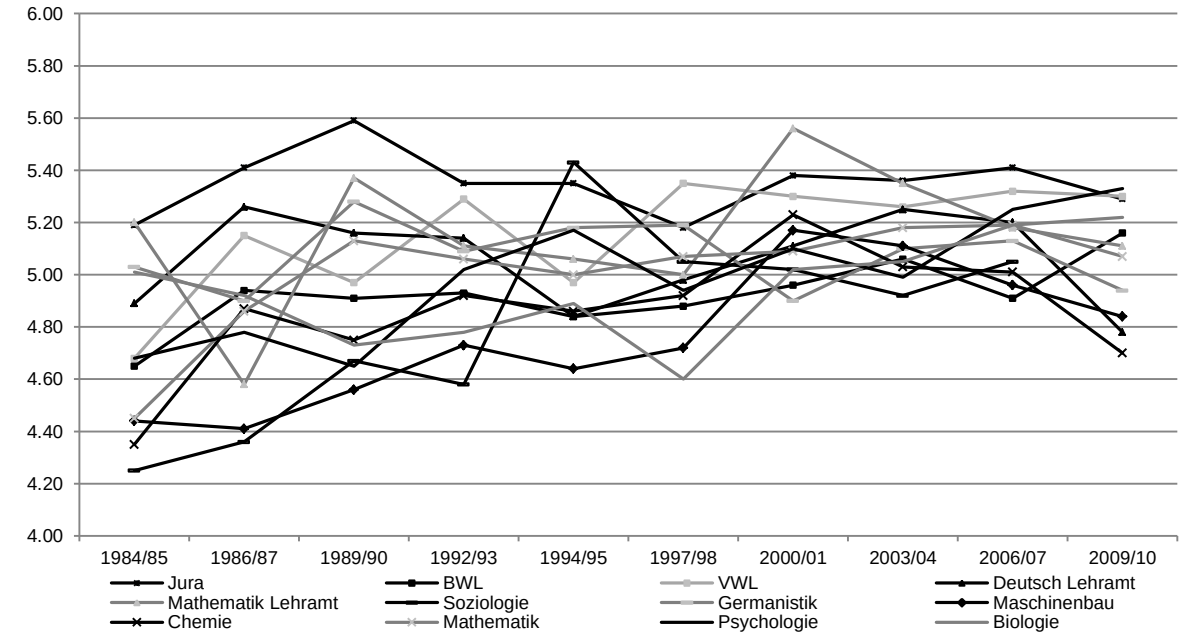
Im Zeitverlauf steigt der durchschnittliche Wert für den Index der beruflichen Stellung in den Fächern/Studiengängen Maschinenbau, Psychologie und VWL sowie in geringerem Umfang in Biologie und BWL über den gesamten Zeitraum tendenziell an. In Mathematik (bis 1990), Soziologie (bis Mitte der 1990er), Mathematik Lehramt und Chemie (beide bis 2000) ist ein phasenweises Ansteigen zu beobachten, in Germanistik Magister, Deutsch Lehramt und Jura ist kein Aufwärtstrend erkennbar.
Studiengangspezifisch ist der notenverbessernde Effekt einer steigenden beruflichen Stellung der Eltern auch wieder nur in Mathematik, Chemie, Biologie, Psychologie, Jura, Germanistik Magister und Deutsch Lehramt zu finden. Da der durchschnittliche berufliche Hintergrund in den letzten drei Studiengängen im Zeitverlauf nicht ansteigt, ist eine Notenverbesserung durch einen gestiegenen beruflichen Hintergrund nur in Mathematik, Chemie, Biologie und Psychologie überhaupt denkbar. Auf Hochschulebene sind keine Daten vorhanden.

Abbildung 126: Anteil Studierende mit mindestens einem akademischen Elternteil



Quelle: Konstanzer Studierendensurvey 1983-2010, eigene Berechnungen,Soziologie inkl. Sozialwissenschaften/Sozialkunde; Maschinenbau inkl. Produktions- und Verfahrenstechnik; Chemie inkl. Bio-/Lebensmittelchemie; Mathematik inkl. Statistik.

Abbildung 127: Durchschnittswert Index der beruflichen Stellung nach Britt Hoffmann (Wertebereich 1-7)



Quelle: Konstanzer Studierendensurvey 1983-2010, eigene Berechnungen, Soziologie inkl. Sozialwissenschaften/Sozialkunde; Maschinenbau inkl. Produktions- und Verfahrenstechnik; Chemie inkl. Bio-/Lebensmittelchemie; Mathematik inkl. Statistik.

Einen im Vergleich zu Teilzeitstudierenden und Stipendiat*innen deutlich höheren Anteil innerhalb der Studierendenschaft weist die Gruppe Studierende mit Migrationshintergrund auf. Laut 20. Sozialerhebung deckt sie (ohne ausländische Studierende) im Sommersemester 2012 23% aller Studierenden ab (Middendorff et al. 2013). Da zumindest für Bildungsinländer gilt, dass sie schlechtere Noten als die übrigen Studierenden erhalten (Grözinger 2015 für die Wirtschaftswissenschaften) und aus dem Schulbereich ebenfalls geringere Bildungserfolge für Migrant*innen bekannt sind (Becker 2011; Diefenbach 2010; Kalter/ Granato 2001; Kristen 2002; 2003), könnten unterschiedliche Anteile

Studierende mit Migrationshintergrund in den Studiengängen wie auch sich verändernde Anteile im Zeitverlauf das durchschnittliche Notenniveau durchaus beeinflussen.

Schlechtere Leistungen von Studierenden mit Migrationshintergrund wären, sollte dieser Zusammenhang existieren, vermutlich vor allem auf Interaktionseffekte zurückzuführen. So weisen allochthone Studierende deutlich häufiger eine niedrige Bildungsherkunft auf, besitzen seltener akademische Elternteile und verfügen auch bei gleichem Bildungshintergrund über weniger finanzielle Absicherung durch die Familie (Middendorff et al. 2013) bei vergleichsweise hohen Bildungsaspirationen (siehe zusammenfassend Becker 2010), was einen besonderen Erfolgsdruck aufbaut. Hinsichtlich des Bildungserfolgs im Schulbereich (Esser 2001; Kalter et al. 2011; Kristen/Granato 2007) und im Hochschulzugang (Kristen et al. 2008) sind jedoch derart deutliche Unterschiede zwischen den verschiedenen in Deutschland lebenden ethnischen Gruppen bekannt, dass ein Versuch, ohne empirische Daten einzig nach dem Kriterium Migrationshintergrund einen möglichen Einfluss auf das durchschnittliche Notenniveau einschätzen zu können, Migrationsforscher*innen die Haare zu Berge stehen lassen würde[87].

Die Entwicklung des Anteils Studierende mit Migrationshintergrund im Zeitverlauf lässt sich nicht einmal auf Fach(gruppen)ebene sinnvoll ermitteln, da zu fehlenden amtlichen Daten auch noch methodische Brüche in der Erfassung des Migrationsstatus in den Sozialerhebungen des Deutschen Studentenwerks kommen. Auch auf Hochschulebene liegen weder im Quer- noch im Längsschnitt brauchbare Daten vor.

Die Geschlechtskomposition würde ebenfalls auf leistungskonforme Weise zu unterschiedlichen Notenniveaus führen, sollten Männer und Frauen unterschiedliche Leistungslevel und damit unterschiedliche Noten im Studium aufweisen. Sollte es einen Unterschied zwischen den Geschlechtern[88] geben, ist davon auszugehen, dass Frauen bessere Noten erzielen als Männer. Entsprechende Befunde lassen sich zum einen bereits seit den 1960er Jahren in Bezug auf Schulnoten finden (vgl. Helbig 2012), zum anderen sind bessere Noten für Frauen in der Zwischenprüfung bekannt (Ramm/Bargel 2005) und kann Grözinger (2015) für die Wirtschaftswissenschaften zeigen, dass im Studium mit steigendem Frauenanteil bessere Durchschnittsnoten einhergehen. Neben einem direkten kompositionellen Effekt ist dabei auch denkbar, dass bei Geschlechtergleichheit ein besseres Lernklima für Frauen herrscht, so dass sie bessere Leistungen produzieren können (Pascarella/Terenzini 2005). Da das

[87] Damit bleibt an dieser Stelle lediglich ein Verweis auf Grözinger (2017), der schlechtere Noten für Bildungsausländer*Innen in allen Fächern findet, die er anhand der Hochschulprüfungsstatistik untersucht. Der Vollständigkeit halber befinden sich zudem fächergruppenspezifische Anteile von Studierenden mit Migrationshintergrund im Anhang (Abb.A28).

[88] Aus Erhebungsgründen kann auch an dieser Stelle nur die bipolare Definition von Geschlecht verwendet werden.

Geschlecht der Absolventinnen und Absolventen in der Datenerhebung mit aufgenommen wurde, lässt sich diese These anhand des Individualdatensatzes (bis 1997) überprüfen.

Eine OLS-Regression der Abschlussnote auf das Geschlecht zeigt über alle Studiengänge hinweg betrachtet und unter Kontrolle des Zeittrends tatsächlich bessere Noten für Frauen. Die Modellgüte ist mit 3% aufgeklärter Varianz allerdings nicht hoch. Zudem liegen möglicherweise studiengangspezifische Verhältnisse vor.

Tabelle 65: OLS-Regression der Abschlussnote auf das Geschlecht 1950-1997

AV: Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Geschlecht männlich[a]	0.070	0.005	14.28	0.000
Jahr	-0.014	0.000	-62.18	0.000
Konstante	29.349	0.439	66.84	0.000
[a]Referenzkategorie: Geschlecht weiblich				n=132 811; r^2adj=0.03

Um ein differenzierteres Bild des Geschlechterverhältnisses zu erhalten, lassen sich die im Durchschnitt erzielten Noten auf Differenzen im Mittelwert testen (Tab.66). Und die Notendifferenz fällt, wie vermutet, je nach Studiengang zugunsten des männlichen oder des weiblichen Geschlechts aus. Diese Erkenntnis alleine ist jedoch nur begrenzt aussagekräftig. Denn soll der Effekt der Notendiskrepanz auf das Gesamtnotenniveau eingeschätzt werden, ist das Verhältnis der Prüflinge entscheidend: Durch die Geschlechterkomposition bedingte gute (schlechte) Durchschnittsnoten bedürfen eines hohen (niedrigen) Anteils des jeweils besseren Geschlechts. Der Effekt der Verteilung der Geschlechter in Kombination mit der festzustellenden Mittelwertdifferenz lässt sich analog zur Verteilung der Akademiker*innenkinder ablesen: Bei einer Gleichverteilung der Geschlechter senkt (hebt) der Notenvorteil (der Notennachteil) einer der Geschlechtergruppen das Gesamtnotenniveau um die Hälfte der Mittelwertdifferenz gegenüber dem Niveau, das bei ebenfalls gleichen Noten von Männern und Frauen herrschen würde (Tab.66, Spalten 3 und 4).

In Germanistik und Mathematik ist die Diskrepanz zwischen den Geschlechtern am größten - die Männer erzielen dort deutlich bessere Noten als die Frauen - in Maschinenbau am geringsten (und auch nicht signifikant). Ohne die tatsächliche Geschlechterverteilung ist der Einfluss der Mittelwertdifferenz auf das Gesamtnotenniveau, wie bereits angesprochen, jedoch noch nicht aussagekräftig - denn: Mit jedem Prozentpunkt, den der Anteil des besser (schlechter) bewerteten Geschlechts steigt (und der des anderen Geschlechts sinkt), verbessert (verschlechtert) sich das Gesamtnotenniveau um den Wert der Mittelwertdifferenz geteilt durch 100 (Spalten 5 und 6).

Dadurch lässt sich nun feststellen, ob die Ungleichverteilung der Geschlechter (bei vorhandener Mittelwertdifferenz) das Gesamtnotenniveau positiv oder negativ beeinflusst. Sowohl in Germanistik als auch in Mathematik erhalten die Männer zwar die besseren Noten, nur in Mathematik sind sie aber auch in der Mehrzahl, was sich im Vergleich zu Germanistik positiv auf das Notenniveau auswirkt - denn dort sind die schlechter bewerteten Frauen in der Überzahl. In Relation der Studiengänge zueinander ergibt sich damit ein (in der Stärke in dieser Reihenfolge abnehmender) notensenkender Ef-

fekt in Mathematik, Chemie, Soziologie Magister, Biologie und Psychologie. In Germanistik, VWL, BWL, Soziologie Diplom, Maschinenbau, Mathematik Lehramt und Deutsch Lehramt ein (ebenfalls in dieser Reihenfolge abnehmender) notenhebender Effekt (Spalte 7).

Unter Verwendung des Aggregatdatensatzes lässt sich der Betrachtungszeitraum erweitern[89]. Spalte 8 in Tabelle 66 gibt an, in wie vielen Jahren der Reihe ein signifikanter Geschlechterunterschied besteht. Der betrachtete Zeitraum beginnt hier in der Regel etwas später als im Vergleich der über die Zeit aggregierten Individualdaten, da beim jahrweisen Vergleich auf die Fallzahl geachtet werden sollte (nur Jahre mit n>2 für beide Geschlechter).

Wie zu erwarten, zeigen sich auch hier studiengangspezifische Resultate: In Soziologie Magister, Maschinenbau und Mathematik Lehramt existieren weder in der Gesamtbetrachtung noch im jahrweisen Vergleich signifikante Unterschiede zwischen den Noten der männlichen und weiblichen Studierenden. Über den gesamten Zeitraum aggregiert, jedoch kaum im Einzelvergleich der Jahre zeigen sich signifikante Unterschiede in VWL und Soziologie Diplom (jeweils bessere Noten für Frauen) und Deutsch Lehramt (bessere Noten für Männer). Hier sind wenige Jahre mit deutlichem Notenunterschied dafür verantwortlich, dass dieser Unterschied auch über den gesamten Zeitraum betrachtet signifikant ist, obwohl unter Berücksichtigung des zeitlichen Aspekts keine nennenswerte Differenz in den Prüfungsergebnissen der Geschlechtergruppen besteht. Über Chemie und Biologie, die immerhin in 15.8% bzw. 18.2% der betrachteten Jahre eine signifikante Differenz (mit besseren Noten für männliche Studierende) aufweisen, führt die Abstufung schließlich zu BWL, Psychologie, Mathematik und Germanistik, die in relevantem zeitlichem Ausmaß signifikante Mittelwertunterschiede in den Noten zwischen den Geschlechtern aufweisen.

Gemäß der dargestellten doppelten Betrachtungsweise kristallisiert sich heraus, dass vor allem diese letztgenannten sechs Studiengänge relevanten Unterschieden in der Notengebung zwischen den Geschlechtern unterliegen. Den größten absoluten Effekt auf die Notendifferenzen zwischen den einzelnen Studiengängen erzielt die unterschiedliche Geschlechterkomposition in Kombination mit diesen Notenunterschieden über den Betrachtungszeitraum bis 1997 dabei im Falle der Differenz zwischen Germanistik (bis 1997 aggregierte Durchschnittsnote: $\bar{x}$=1.94) und Mathematik ($\bar{x}$=1.67). Im Vergleich zur Differenz, die das Gesamtnotenniveau zwischen diesen beiden Studiengängen aufweisen würde, wenn die Anteile und Noten beider Geschlechter gleich wären, ergibt sich eine *maximale*[90] Vergrößerung von 0.227 (0.119 Verbesserung in Mathematik + 0.108 Verschlechterung in Ger-

[89] Für Soziologie stehen nach 1997 keine geschlechtsspezifischen Informationen zur Verfügung.

[90] Analog zu den Veränderungswerten für den Einfluss des Anteils Akademiker*innenkinder gilt hier, dass die maximale Veränderung gegenüber der Annahme gleicher Geschlechteranteile mit gleichen Durchschnittsnoten nur erreicht wird, wenn sich die Veränderung, die sich bei gleich hohen Anteilen durch die Mittelwertdifferenz

manistik) Noten. Dies sind immerhin 85.3 % der Differenz von 0.266 zwischen den aggregierten Durchschnittsnoten. Werden Mathematik und Chemie zum Vergleich untereinander herangezogen, wird deutlich, dass auch hier nur die *maximal mögliche* Veränderung durch *ein einzelnes* Zusammensetzungsmerkmal berechnet wird: Die Veränderungswerte stellen 180.9% der tatsächlich vorhandenen Differenz im Notenniveau dar.
Wie stark studiengangabhängig die Rechnung ist, zeigt sich auch in den restlichen Vergleichen. In folgender Reihenfolge verringert sich das maximale Erklärungspotential der Notendifferenzen zwischen den Paaren: Germanistik-Chemie (64.8%), Germanistik-Psychologie (29.7%), Germanistik-Biologie (24.8%), BWL-Mathematik (14.8%), Biologie-Psychologie (7.7%), BWL-Chemie (7.6%), BWL-Biologie (5.1%), BWL-Psychologie (4.8%). Für die Paare Germanistik-BWL, Mathematik-Psychologie, Mathematik-Biologie, Psychologie-Chemie und Biologie-Chemie tragen die Geschlechterdisparitäten definitiv nicht zur Erklärung der Differenzen bei - hier würde sich die Differenz im Notenniveau sogar noch vergrößern, wären die Anteile und Noten jeweils gleich verteilt.

Hinsichtlich der langfristigen Notenentwicklung findet sich die Annahme, die Durchschnittsnoten hätten sich in den vergangenen Jahrzehnten durch einen steigenden Anteil an Frauen verbessert. Wie in Abbildung 128 ersichtlich, nimmt der Frauenanteil über die Zeit in allen Studiengängen zu, allerdings in deutlich unterschiedlicher Stärke und Geschwindigkeit: In Germanistik sind die Frauen bereits zu Beginn der Reihe 1970 das dominante Geschlecht und bleiben dies auch bis auf eine kurze Phase Mitte der 1970er, auf die nach kurzem Absinken ein relativ konstantes Anwachsen des Anteils bis zum Ende der Zeitreihe folgt. In Psychologie liegt der Frauenanteil bis zum Ende der 1970er Jahre knapp unter 50%, danach folgen ein starker Anstieg und eine Dominanz der weiblichen Studierenden. In Deutsch Lehramt ist ein zu Beginn der 1990er Jahre abflachender Anstieg des Frauenanteils zu beobachten, der dazu führt, dass der anfängliche Männerüberschuss ebenfalls in der Mitte der 1970er in einen Frauenüberschuss umschlägt. In Biologie findet sich eine konstante Zunahme, der Männerüberschuss ist Ende der 1980er Jahre ausgeglichen. In Mathematik Lehramt beginnt der Frauenanteil erst zu Beginn der 1970er sichtbar zuzunehmen und nach einem Hoch von ca. 40% Anfang der 1990er sinkt er 10 Jahre, bevor er zum Ende der Reihe hin wieder ansteigt. In Soziologie mit Abschluss Magister sind die weiblichen Studierenden seit Mitte der 1990er in der Überzahl, nachdem dort ein langsameres, ab Mitte der 1980er dann starkes, zyklisches Anwachsen des Frauenanteils zu verzeichnen ist.

---

alleine ergibt das Notenniveau in die gleiche Richtung verändert, in die auch die unterschiedlich hohen Anteile bei gegebener Differenz wirken.

Tabelle 66: Studiengangspezifische Mittelwertvergleiche (T-Test) der Noten zwischen den Geschlechtergruppen und Auswirkungen auf das Notenniveau

Studiengang	Zeitraum	Aggregierte Mittelwertdifferenz[a]	Veränderung gegenüber gleicher Durchschnittsnote	Anteil weiblicher Prüflinge	Veränderung gegenüber gleichem Anteil	Maximale Veränderung	Anzahl auf 5% Niveau signifikanter Jahre[b] (abweichender Zeitraum)
Germanistik	1962-1997	-0.170**	±0.085	63.6%	+0.023	+0.108	19/45 (1966-2010)
Mathematik	1950-1997	-0.145***	±0.073	18.2%	-0.046	-0.119	12/52 (1959-2010)
Psychologie	1951-1997	0.025*	±0.013	57.2%	-0.002	-0.014	12/52 (1959-2010)
BWL	1953-1997	0.060***	±0.030	22.5%	+0.017	+0.047	10/53 (1957-2009)
Biologie	1951-1997	-0.045***	±0.023	49.8%	-0.000	-0.023	8/44 (1967-2010)
Chemie	1950-1997	-0.041*	±0.021	16.0%	-0.014	-0.034	9/57 (1950-2008)
Deutsch Lehramt	1958-1997	-0.041***	±0.021	57.4%	+0.003	+0.017	6/48 (1963-2010)
VWL	1950-1997	0.061***	±0.031	21.4%	+0.017	+0.048	5/61 (1950-2010)
Mathematik Lehramt	1959-1997	0.027	±0.014	32.6%	+0.005	+0.018	2/50 (1961-2010)
Maschinenbau	1960-1997	0.024	±0.012	1.9%	+0.012	+0.024	1/35 (1976-2010)
Soziologie Diplom	1962-1997	0.048*[c]	±0.024	45.4%	+0.002	+0.026	1/36 (1962-1997)
Soziologie Magister	1963-1997	-0.048	±0.024	43.6%	-0.003	-0.027	0/29 (1969-1997)

*p≤0.05 **p≤0.01 ***p≤0.001

[a]Mittelwertdifferenz=Durchschnittliche Abschlussnote männlich – Durchschnittliche Abschlussnote weiblich

[b]nur Jahre mit n>2 für beide Geschlechter berücksichtigt

[c]nur FU Berlin

In den beiden wirtschaftswissenschaftlichen Studiengängen ist zwar ebenfalls ein starkes (zyklisches) Wachstum des Anteils der weiblichen Studierenden ab Anfang/Mitte der 1970er Jahre festzustellen, aber auch am Ende der Reihe herrscht dort noch Männerüberschuss. Ähnlich verhält es sich in Mathematik und Chemie, wo das Wachstum zudem etwas schwächer ausfällt als etwa in BWL. Die geringste Bedeutung hat der Frauenzuwachs in Maschinenbau: Dort ist nur ein sehr schwaches Wachstum zu beobachten, das ab Anfang 1990er ein wenig stärker wird, an dessen Ende der Frauenanteil aber immer noch bei gerade einmal knapp über 10% liegt.

Grundsätzlich ist eine Beeinflussung der Notendurchschnitte im Zeitverlauf durch mehr weibliche Studierende für einzelne Studiengänge also eine plausible Annahme - aber nur möglich, wenn Frauen dort auch bessere Noten erzielen: Erzielt eines der beiden Geschlechter bessere Noten als das andere, verbessert (verschlechtert) sich mit steigendem (sinkenden) Anteil dieser Gruppe auch das mittlere Notenniveau. Wie sich gezeigt hat, existieren durchaus Geschlechterunterschiede, die sich allerdings studiengangspezifisch darstellen. Auch im Zeitverlauf muss das Geschlechterverhältnis nach Studiengängen betrachtet werden, wie Abb.129 verdeutlicht. Sie zeigt, dass sich über alle Studiengänge des samples betrachtet für die meisten Jahre bessere Noten für Männer finden lassen. In Mathematik, Deutsch Lehramt und Germanistik erzielen sie durchgehend bessere Ergebnisse. Allerdings sind in einigen Studiengängen phasenweise auch bessere Noten für Frauen zu beobachten: In BWL ab Mitte/Ende der 1980er sowie in VWL und Chemie ab Beginn der 1990er Jahre. In Psychologie ist kein Trend erkennbar, das Notenniveau ist relativ ausgeglichen, mit einem durchgängig leichten Vorteil für Frauen, die in Soziologie nur in wenigen Jahren Anfang der 1990er, in Biologie in einigen Jahren zum Ende der 2000er besser benotet werden. In Mathematik Lehramt erhalten weibliche Studierende zwischen Mitte der 1960er und Mitte der 1970er leicht bessere Noten, bevor sich die Differenz ausgleicht und ab Mitte der 1990er die Männer die Rolle der etwas besser bewerteten Prüflinge übernehmen. In Maschinenbau liegt die absolute Anzahl an weiblichen Studierenden erst ab 1990 nicht mehr unter n=10, der Zeitraum vorher erlaubt keine sinnvolle Interpretation der Zahlen. Mit dem ersten Anwachsen des Frauenanteils geht eine kurze Phase besserer Noten für Frauen einher, bevor sich leicht bessere Noten für Männer einstellen.

Der wachsende Anteil der weiblichen Studierenden kann damit überhaupt nur in vier Studiengängen und auch dort nur zeitlich begrenzt zu einer spürbaren Verbesserung des Notenniveaus geführt haben: In BWL ab Mitte/Ende der 1980er Jahre sowie in VWL und Chemie ab Beginn der 1990er Jahre, seit dort die Frauen parallel zu ihrem anteilsmäßigen Zugewinn auch bessere Noten erzielen, und in Psychologie ab Mitte der 1970er Jahre, seit dort der Anteil der Frauen, die im Studiengang durchgehend leicht bessere Noten erzielen, wächst. Wie stark der Effekt des wachsenden Frauenanteils bei besseren Noten ist, lässt sich anhand der jährlichen Veränderung des prozentualen Anteils und der in den jeweiligen Jahren herrschenden Notendifferenz zwischen den Geschlechtern berechnen. Nur für

VWL und Psychologie ergibt sich eine nennenswerte Veränderung: Im Zeitraum der maximalen Verbesserung durch geschlechtsspezifische Notenniveaus in Kombination mit Veränderungen im Frauenanteil ergibt sich für VWL zwischen 1991 und 2002 ein niveausenkender Einfluss von 0.11 Noten, in Psychologie zwischen 1969 und 2010 von 0.17 Noten. Im jährlichen Durchschnitt beträgt der verbessernde Einfluss damit in VWL 0.010, in Psychologie 0.004 Noten. In Chemie (1991-2006: Verbesserung um 0.08 Noten=0.005 Noten/Jahr) und BWL (1984-2009: Verbesserung um 0.03 Noten=0.001 Noten/Jahr) ist dieser Effekt deutlich geringer.

Abbildung 128: Prozentualer Anteil der weiblichen Studierenden nach Studiengang im Zeitverlauf (LOWESS 0.3)

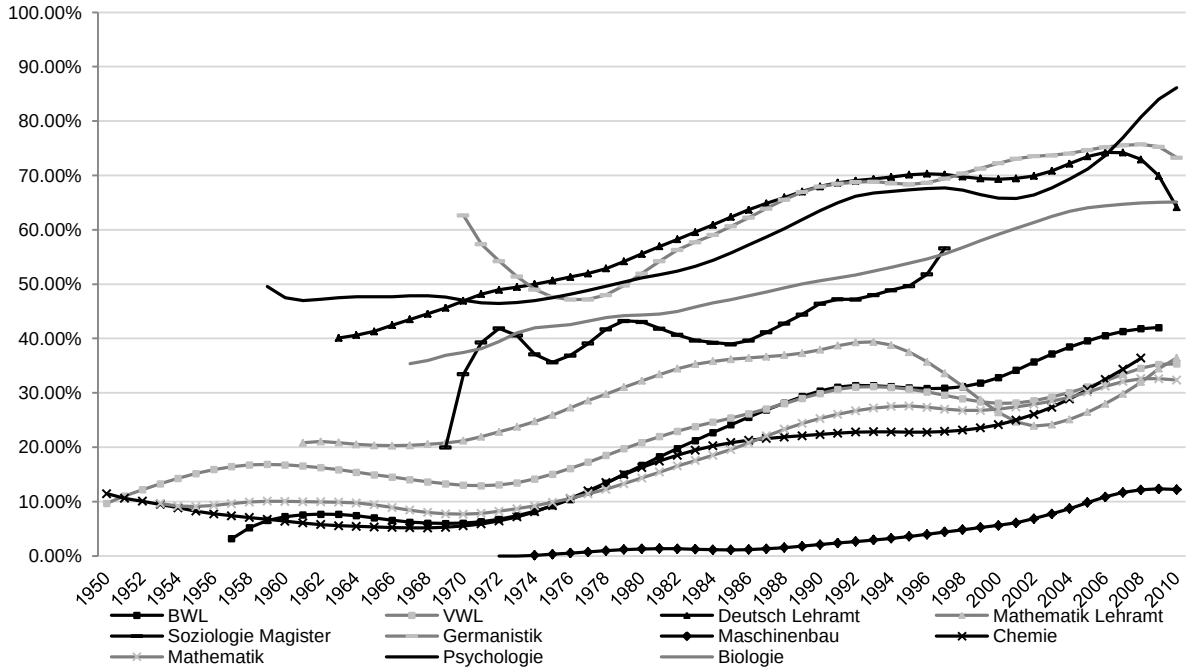
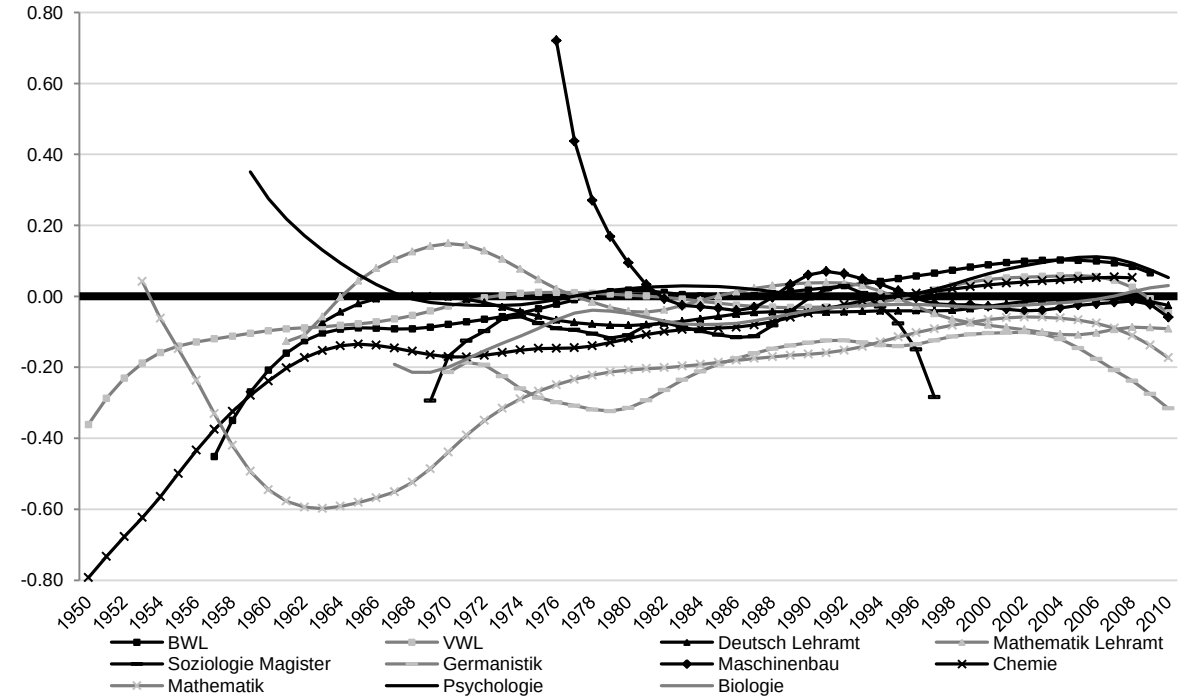


Abbildung 129: Differenz durchschnittliche Abschlussnote der männlichen - der weiblichen Studierenden (LOWESS 0.4)



Lesehilfe: Werte im positiven Bereich bedeuten im Durchschnitt bessere, Werte im negativen Bereich im Durchschnitt schlechtere Noten für weibliche Studierende gegenüber männlichen Studierenden.

Auch auf Hochschulebene liegen Daten zur Überprüfung der Geschlechterthese vor, hier allerdings nur die Archivdaten, da das Forschungsdatenzentrum auf Hochschulebene unter Verweis auf Datenschutzbedenken für zu wenige Jahre Werte zur Verfügung stellt, als dass sich die Zeitreihen bis 2010 weiterführen ließen[91]. Hier offenbaren Mittelwertunterschiede zwischen den Geschlechtern je nach Studiengang variierende Unterschiede zwischen den Standorten (Tab.67).

Im Diplom Mathematik weisen Männer an allen Hochschulen des samples bessere Noten auf, allerdings in unterschiedlichem Ausmaß. In Berlin und Göttingen liegt der Notenvorteil der Männer um ca. 0.15 Noten höher als in Heidelberg, Karlsruhe und Braunschweig (Spalte 4). In Münster und Tübingen ist im Vergleich nur ein minimaler Vorteil gegeben. Die Signifikanzwerte sind an dieser Stelle von geringerer Bedeutung als üblich, da es sich bei den Daten auf Hochschulebene um eine Vollerhebung handelt - unabhängig davon, ob die Mittelwertunterschiede signifikant sind oder nicht, sie existieren auf diesem Aggregatlevel in exakt dieser Größe.

Da die berechneten Notenunterschiede über einen langen Zeitraum aggregiert berechnet wurden, sollte überprüft werden, ob es sich dabei auch um langfristig konstante Unterschiede handelt oder ob sie im Aggregat aufgrund kurzer Perioden mit starken Differenzen zustande kommen. In Spalte 9 ist aufgeführt, in wie vielen Jahren Männer an den einzelnen Hochschulen bessere Noten als Frauen erzielt haben. Dabei wurden nur Jahre betrachtet, die für beide Geschlechter n>2 aufweisen. Entsprechend dem Ausmaß der Differenz ist der Anteil der Jahre, in denen die Männer bessere Noten erzielen in Berlin und Göttingen am höchsten.

Wird der Beobachtungszeitraum angeglichen (1971-1997), verringern sich die Vorteile in Göttingen und Berlin um 0.04 bzw. 0.05 Noten, liegen damit aber immer noch über denen der anderen Hochschulen. Damit übt der Geschlechteranteil einen merkbaren Einfluss auf die Notenunterschiede zwischen Göttingen/Berlin auf der einen und Münster/Tübingen auf der anderen Seite aus. Da die Noten in Göttingen und Münster jedoch tendenziell über denen in Berlin und Tübingen liegen, ist die Erklärungskraft für die aufgezeigten Notenunterschiede jedoch begrenzt.

Im Lehramtsstudiengang Mathematik zeigen die Vorzeichen der Mittelwertdifferenzen zwar in unterschiedliche Richtungen, allerdings sind die Differenzen gering und liegen aggregiert allesamt unter 0.07 Noten. Nur Braunschweig fällt mit einem deutlichen Notenvorteil von 0.210 Noten zugunsten der Frauen aus der Rolle. Entsprechend gering ist der Anteil Jahre, in denen Männer die besseren Noten aufweisen. Auffällig ist hierbei, dass Braunschweig sowohl die geringste Gesamtzahl Studie-

[91] Die Zeitreihen verkürzen sich bei geschlechterspezifischer Betrachtung der Noten zudem dadurch, dass zu Beginn der Reihen in den meisten Studiengängen nur eine sehr geringe Anzahl an Frauen geprüft wurde. Um die starken Schwankungen auszuschließen, die bei der Mittelwertbildung durch die geringe Fallzahl entstehen, wurden die Jahre zu Beginn der Zeitreihen aus der Analyse ausgeschlossen die für eines der beiden Geschlechter ein n<3 aufweisen (Ausnahme: Es handelt sich nur um ein einziges Jahr zu Beginn der Reihe).

rende in Mathematik Lehramt aufweist, als auch den geringsten Anteil Frauen. Entsprechend kann die Geschlechterkomposition auch nur im Vergleich der TU mit den anderen Standorten einen Einfluss auf Unterschiede im Notenniveau besitzen.

In Chemie ist der Einfluss der Geschlechterkomposition auf die Notendifferenzen vergleichsweise niedrig einzuschätzen. Nur in Münster weisen Frauen die besseren Noten auf. In Tübingen, Heidelberg und Karlsruhe ist der Notenvorteil der Männer am größten. Es ergibt sich damit im Vergleich zur Differenz, die das Gesamtnotenniveau jeweils aufweisen würde, wenn die Anteile und Noten beider Geschlechter gleich wären, eine maximale absolute Veränderung von 0.188 Noten zwischen Tübingen (Verbesserung von 0.133 Noten) und Münster (Verschlechterung von 0.055 Noten). In 15 von 21 Paarvergleichen wären die Differenzen unter Annahme einer ausgeglichenen Verteilung und gleicher Durchschnittsnoten größer als unter den tatsächlichen Verhältnissen. Nur für die sechs Paare Münster-Berlin (87.9%), Münster-Braunschweig (28.5%), Münster-Heidelberg (22.4%), Karlsruhe-Tübingen (20.9%), Münster-Göttingen (18.0%) und Karlsruhe-Heidelberg (5.0%) ergäbe sich unter den modellhaften Annahmen eine Verringerung der vorzufindenden Niveauunterschiede.

In Biologie schneiden über den jeweils gesamten Beobachtungszeitraum stets die Männer besser ab, allerdings beträgt ihr Vorteil gegenüber den Frauen maximal 0.063 Noten (in Berlin). Bei überall fast ausgeglichenen Geschlechteranteilen ergibt sich damit kaum ein Einfluss auf die festgestellten Unterschiede im Notenniveau. In Psychologie fallen die Mittelwertdifferenzen ebenfalls gering aus. Durch unterschiedliche Vorzeichenrichtungen erhöhen sich die Diskrepanzen zwischen den Standorten zwar im Vergleich zu Biologie, allerdings sind auch hier die Geschlechteranteile relativ ausgeglichen, so dass sich kein großer maximaler Einfluss auf das Notenniveau ergibt.

In VWL erzielen mit Ausnahme von Tübingen überall die Frauen die besseren Noten, wobei Karlsruhe einen deutlichen Vorteil der weiblichen Studierenden aufweist. Im Gegensatz zum Ausreißer Braunschweig im Studiengang Mathematik Lehramt ist hier allerdings der höchste Anteil Frauen zu verzeichnen. Gemeinsam ist beiden jedoch, dass die Gesamtzahl der Fälle deutlich niedriger liegt als an den anderen Hochschulen. Im Paarvergleich ist die Geschlechterkomposition also nur unter Beteiligung von Karlsruhe und Tübingen relevant – zwischen diesen beiden ergibt sich eine maximale Verringerung der festgestellten Differenz im Notenniveau um fast eine Drittelnote. In BWL halten sich die Paarvergleiche mit einer Vergrößerung und die mit einer Verringerung der Unterschiede die Waage. Für jeweils drei Paare (Tübingen-Göttingen mit 18.1%, Tübingen-Münster mit 17.6% und Münster-Göttingen mit 13.2% der Differenz) lässt sich ein Teil der Differenzen durch die ungleichen Geschlechterverhältnisse erklären, für die anderen drei erhöht sich der zu erklärende Unterschied.

In Soziologie ist der Notenvorteil wiederum standortspezifisch - in Münster und Göttingen fällt die Differenz zugunsten der männlichen, in Tübingen, Berlin (Diplom) und Heidelberg zugunsten der weiblichen Absolvent*innen aus. Göttingen weicht mit einem Wert von 0.165 Noten deutlich von den geringen Unterschieden an den anderen Hochschulen ab, durch die Gleichverteilung der Geschlechter dort bleibt der maximale Einfluss auf das Notenniveau jedoch wie auch überall anderorts unter einer Zehntelnote. Der Anteil Jahre, in denen die Männer bessere Noten erzielen, liegt an allen Standorten um die 50% herum und zeugt damit auch von etwa ausgeglichenen Verhältnissen.

In Germanistik sind in der aggregierten Betrachtung an allen Hochschulen bessere Noten für Männer zu finden. in Münster fällt der Notenvorteil mit 0.226 Noten knapp dreieinhalbmal so hoch aus wie in Heidelberg mit 0.066 Noten. In Kombination mit ebenfalls vorhandenen Unterschieden im Frauenanteil, der eine Spannweite von 11 Prozentpunkten aufweist, ergibt sich durch die Geschlechterkomposition eine maximale Verschlechterung gegenüber auf die Geschlechter gleich verteilten Anteilen und Noten zwischen 0.044 Noten in Heidelberg und 0.154 Noten in Münster.
Für die Differenzen im Notenniveau zwischen den einzelnen Standorten ergibt sich eine maximale absolute Veränderung von 0.110 (0.154 Verschlechterung in Münster - 0.044 Verschlechterung in Heidelberg) Noten. Diese Differenz beträgt 102.8% der tatsächlichen Differenz zwischen den beiden Universitäten, welche sich damit vollständig durch die Geschlechterkomposition erklären ließe, angenommen, die Noten unterlägen nur diesem Einflussfaktor. In folgender Reihenfolge sinkt der Anteil der maximalen Erklärungskraft an den Notendifferenzen: Tübingen-Saarbrücken (58.7%), Münster-Saarbrücken (55.0%), Münster-Tübingen (52.8%), Münster-Berlin (18.5%), Göttingen-Heidelberg (13.9%), Göttingen-Saarbrücken (7.6%), Berlin-Tübingen (6.7%), Berlin-Göttingen (2.2%).
Für die übrigen sechs Paarvergleiche würde eine ausgeglichene Geschlechterkomposition mit gleichen Noten die Abstände im Notenniveau unter den gegebenen Verhältnissen im Gegenteil vergrößern. Die Männer sind an allen Hochschulen in mindestens der Hälfte aller Jahre im Vorteil, die anteilsmäßig größere Gruppe der schlechter bewerteten Frauen führt dort also nicht nur punktuell zu Verschlechterungen des Notenniveaus

Auch in Deutsch Lehramt weisen die Männer in der Regel die besseren Durchschnittsnoten auf, nur in Göttingen fällt die Mittelwertdifferenz zugunsten der Frauen aus. Da die Anteile der Geschlechter relativ ausgeglichen sind, liegen die maximalen Veränderungswerte des Notenniveaus unter einer Zehntelnote. Und auch bei gegensätzlichem Vorzeichen von Göttingen und den übrigen Hochschulen bleibt die Spannweite der maximalen Veränderung im Paarvergleich stets unter einer Zehntelnote.

Tabelle 67: Hochschulspezifische Mittelwertvergleiche (T-Test) der Noten zwischen den Geschlechtergruppen und Auswirkungen auf das Notenniveau

Studiengang	Hochschule	Zeitraum	Aggregierte Mittelwertdifferenz[a] (Zeitraum)	Veränderung gegenüber gleicher Durchschnittsnote	Anteil weiblicher Prüflinge	Veränderung gegenüber gleichem Anteil	Maximale Veränderung	Anzahl Jahre mit Notenvorteil männlicher Prüflinge[b] (Zeitraum)
Mathematik D	Göttingen	1950-1997	-0.260***	±0.130	16.0%	-0.088	-0.218	21/25 (1968-1997)
	Berlin	1953-1997	-0.299***	±0.150	16.0%	-0.102	-0.252	25/31 (1954-1997)
	Karlsruhe	1960-1997	-0.116*	±0.058	20.0%	-0.035	-0.093	13/19 (1973-1997)
	Münster	1966-1997	-0.077	±0.039	19.5%	-0.023	-0.062	14/21 (1974-1997)
	Heidelberg	1964-1997	-0.136*	±0.068	14.4%	-0.048	-0.116	13/21 (1971-1997)
	Braunschweig	1971-1997	-0.114	±0.057	26.0%	-0.027	-0.084	13/21 (1973-1997)
	Tübingen	1963-1997	-0.020	±0.010	17.0%	-0.007	-0.017	9/17 (1967-1997)
MathematikLA	Tübingen	1973-1997	-0.029	±0.015	36.2%	-0.004	-0.019	15/25 (1973-1997)
	Berlin	1963-1997	0.038	±0.019	31.8%	+0.007	+0.026	17/34 (1963-1997)
	Karlsruhe	1963-1997	-0.036	±0.018	33.9%	-0.006	-0.024	16/34 (1964-1997)
	Braunschweig	1961-1997	0.210***	±0.105	26.0%	+0.050	+0.155	9/25 (1961-1997)
	Göttingen	1959-1997	0.067	±0.034	31.3%	+0.013	+0.047	11/32 (1960-1997)
Chemie	Karlsruhe	1960-1997	-0.114*	±0.057	15.9%	-0.039	-0.096	21/26 (1960-1997)
	Göttingen	1950-1997	-0.002	±0.001	16.2%	-0.001	-0.002	19/26 (1950-1997)
	Tübingen	1970-1997	-0.159**	±0.080	16.7%	-0.053	-0.133	16/23 (1974-1997)
	Heidelberg	1959-1997	-0.130***	±0.065	12.7%	-0.048	-0.113	21/33 (1960-1997)
	Braunschweig	1970-1997	-0.029	±0.015	21.5%	-0.008	-0.023	15/24 (1973-1997)
	Münster	1950-1997	0.065	±0.033	16.1%	+0.022	+0.055	15/28 (1950-1997)
	Berlin	1950-1997	-0.055	±0.028	16.2%	-0.019	-0.047	13/28 (1955-1997)
Biologie	Münster	1976-1997	-0.034	±0.017	49.9%	-0.000	-0.017	14/20 (1976-1997)
	Göttingen	1981-1997	-0.053*	±0.027	50.4%	+0.000	+0.027	11/16 (1982-1997)
	Tübingen	1967-1997	-0.043*	±0.022	47.6%	-0.001	-0.023	21/31 (1967-1997)
	Berlin	1951-1997	-0.063**	±0.032	51.2%	+0.001	+0.033	19/29 (1954-1997)
	Heidelberg	1970-1997	-0.046	±0.023	48.7%	-0.001	-0.024	16/27 (1971-1997)
	Karlsruhe	1982-1997	-0.031	±0.016	57.0%	+0.002	+0.018	8/14 (1984-1997)
	Braunschweig	1971-1997	-0.003	±0.002	49.9%	-0.000	-0.002	13/25 (1972-1997)
Psychologie	Göttingen	1967-1997	-0.065	±0.033	60.0%	+0.007	+0.040	23/31 (1967-1997)
	Berlin	1971-1997	-0.040*	±0.020	62.7%	+0.005	+0.025	15/27 (1971-1997)
	Tübingen	1959-1997	0.035	±0.018	51.0%	-0.000	-0.018	21/38 (1959-1997)
	Heidelberg	1960-1997	0.027	±0.014	57.3%	-0.002	-0.016	14/34 (1960-1997)
	Braunschweig	1972-1997	0.041	±0.022	55.3%	-0.002	-0.024	10/26 (1972-1997)
	Münster	1951-1997	0.023	±0.012	54.3%	-0.001	-0.013	13/34 (1964-1997)

noch Tabelle 67: Hochschulspezifische Mittelwertvergleiche (T-Test) der Noten zwischen den Geschlechtergruppen und Auswirkungen auf das Notenniveau

	Münster	**1950-1997**	**0.024**	**±0.012**	**20.0%**	**+0.007**	**+0.019**	**33/47 (1950-1997)**
	Tübingen	1950-1997	-0.084*	±0.042	16.2%	-0.028	-0.070	29/45 (1950-1997)
VWL	Göttingen	1963-1996	0.041	±0.022	20.9%	+0.012	+0.034	19/31 (1964-1996)
	Heidelberg	1955-1997	0.069*	±0.035	26.1%	+0.016	+0.051	22/38 (1956-1997)
	Berlin	1953-1997	0.045	±0.023	21.4%	+0.013	+0.036	20/45 (1953-1997)
	Karlsruhe	1960-1997	0.286***	±0.143	26.3%	+0.068	+0.211	7/24 (1974-1997)
	Tübingen	1984-1997	-0.008	±0.004	30.7%	-0.015	-0.019	8/12 (1986-1997)
BWL	Münster	1957-1997	0.047***	±0.024	21.6%	+0.013	+0.037	23/39 (1959-1997)
	Berlin	1953-1997	0.112***	±0.056	22.5%	+0.031	+0.087	19/44 (1953-1997)
	Göttingen	1963-1996	0.042*	±0.021	23.0%	+0.011	+0.032	14/33 (1963-1996)
	Münster	1967-1997	-0.036	±0.018	39.6%	-0.004	-0.022	10/18 (1972-1997)
	Göttingen	1969-1997	-0.165*	±0.083	50.6%	+0.000	+0.083	13/24 (1971-1997)
Soziologie	Tübingen	1963-1997	0.055	±0.028	37.3%	+0.007	+0.035	10/20 (1965-1997)
	Berlin(Diplom)	1962-1997	0.048*	±0.024	45.4%	+0.002	+0.026	16/35 (1963-1997)
	Heidelberg	1970-1997	0.022	±0.011	47.8%	+0.000	+0.011	10/22 (1970-1997)
	Berlin	1970-1997	-0.127***	±0.064	59.5%	+0.012	+0.076	24/27 (1971-1997)
	Münster	1962-1997	-0.226***	±0.113	68.0%	+0.041	+0.154	14/20 (1974-1997)
Germanistik	Göttingen	1969-1997	-0.127*	±0.064	70.5%	+0.026	+0.090	14/20 (1978-1997)
	Tübingen	1964-1997	-0.155**	±0.078	62.5%	+0.019	+0.097	14/21 (1971-1997)
	Heidelberg	1970-1997	-0.066	±0.033	66.8%	+0.011	+0.044	18/27 (1971-1997)
	Saarbrücken	1970-1997	-0.092	±0.046	65.2%	+0.014	+0.060	6/12 (1983-1997)
	Karlsruhe	1963-1997	-0.089***	±0.045	60.8%	+0.010	+0.055	28/35 (1963-1997)
	Tübingen	1973-1997	-0.124***	±0.062	56.4%	+0.008	+0.070	19/25 (1973-1997)
DeutschLA	Braunschweig	1972-1997	-0.098	±0.049	55.7%	+0.006	+0.055	16/24 (1974-1997)
	Göttingen	1958-1997	0.054*	±0.027	53.5%	-0.002	-0.029	24/39 (1959-1997)
	Berlin	1963-1997	-0.010	±0.005	58.4%	+0.001	+0.006	19/35 (1963-1997)

*p≤0.05 **p≤0.01 ***p≤0.001

[a]Mittelwertdifferenz=Durchschnittliche Abschlussnote männlich – Durchschnittliche Abschlussnote weiblich

[b]nur Jahre mit n>2 für beide Geschlechter berücksichtigt

Im Zeitverlauf gilt auch auf Hochschulebene dass ein Anstieg der Anteile der weiblichen Studierenden nur dann zu einem Absinken des Durchschnittsnotenniveaus geführt haben kann, wenn Frauen in den Perioden, in denen ihr Anteil steigt, bessere Noten erzielen als Männer. Wie nach der Analyse auf Studiengangebene nicht anders zu erwarten, ist dies nur selten der Fall. Nur 13 von 56 Zeitreihen, die ausreichend weibliche Fallzahlen bieten, um sie über einen längeren Zeitraum zu betrachten, weisen überhaupt eine ungefähre Parallelität der beiden Bedingungen auf. Nur zweimal (Chemie und BWL, jeweils in Berlin) treffen beide Bedingungen als kontinuierliche Entwicklung auf und beschränken sich nicht auf einen ca. 10 Jahre andauernden Zeitraum im Verlauf der Reihen, wie in den übrigen 11 Betrachtungen[92]. Im Zeitraum der maximalen Verbesserung durch geschlechtsspezifische Notenniveaus in Kombination mit Veränderungen im Frauenanteil ergibt sich an der Universität Berlin für Chemie zwischen 1985 und 1997 ein niveausenkender Einfluss von 0.80 Noten, für BWL zwischen 1980 und 1995 von 0.41 Noten. Im jährlichen Durchschnitt beträgt der verbessernde Einfluss damit in Chemie 0.067, in BWL immerhin noch 0.028 Noten. In diesen beiden Studiengängen ist in diesen Phasen damit ein erheblicher Einfluss der Geschlechtskomposition zu verzeichnen.

Es zeigt sich, dass die Geschlechtskomposition Auswirkungen auf das Notenniveau haben *kann*. Im aggregierten Querschnitt, der über so lange Zeiträume wie hier zusammengefasst wird, sind die Resultate vor allem als Anhaltspunkte zu verstehen, die Tendenzen wiedergeben können, aber keine genauen Effektstärken. Ob die konkrete Geschlechterverteilung in Kombination mit den gegebenen Notenvorteilen einer der beiden Gruppen zur Erklärung der festgestellten Niveauunterschiede in den Abschlussnoten beiträgt oder sie sogar noch vergrößert, ist abhängig vom betrachteten Paarvergleich. Ein Einfluss des Geschlechts auf die Notenhöhe ist zwischen einzelnen Studiengängen durchaus gegeben, kann aber keinesfalls als systematisch eingestuft werden.
Auf Hochschulebene innerhalb der Studiengänge ist kein Muster erkennbar. Im Längsschnitt zeigt sich auf Studiengangebene nur in VWL und Psychologie ein nennenswerter Effekt über einen Zeitraum von circa einer Dekade. Insgesamt ist der Anstieg des Anteils der weiblichen Studierenden damit keine starke und vor allem keine systematische Ursache für Notenverbesserung. Und auch auf Hochschulebene ist die Entwicklung der Geschlechtskomposition nur in zwei Fällen - dort allerdings im entsprechenden Zeitraum in nennenswertem Ausmaß - relevant.

[92] Auch ein umgekehrter Mechanismus, nach dem sinkende Frauenanteile bei (häufiger als umgekehrt) auftretenden Notenvorteilen für Männer die Notengebung zum Besseren beeinflussen ist im Zeitraum bis 1997 nicht nachweisbar, da ein Absinken des Anteils weiblicher Prüflinge nur sehr selten, und wenn dann frühestens zu Beginn der 1990er Jahre, auftritt.

Eingangseignung

Wie bereits in Kapitel 6.2 beschrieben, muss als kompositioneller Einflussfaktor auf die Notengebung auch die (von soziodemographischen Merkmalen unabhängige?) durchschnittliche Eingangseignung der Studierenden berücksichtigt werden. Dem inzwischen eindeutig belegten Befund, dass bessere Abiturnoten mit größerem Studienerfolg einhergehen (Trapmann et al. 2007; zusammenfassend: Köller 2013) entsprechend, zeigt sich auch hinsichtlich des Zwischenprüfungsergebnisses ein fachübergreifender Einfluss der Eingangseignung: Je besser die Abiturnote, umso besser ist auch die Note der Befragten der im sample enthaltenen Fächer/Studiengänge im Konstanzer Studierendensurvey. Dieser Zusammenhang ist auch für jedes Fach/Studiengang separat berechnet hochsignifikant.

Tabelle 68: OLS-Regression der Zwischenprüfungsnote auf die Abiturnote

AV: Zwischenprüfungsnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Abiturnote	0.393	0.010	38.32	0.000
Konstante	1.703	0.024	70.22	0.000
				n=9535; r^2=0.13

Durch unterschiedliche Kompetenzniveaus, die durch einen Selektionsprozess der begabteren bzw. stärker motivierten Studierenden in bestimmte Fächer bzw. Studiengänge oder an bestimmte Hochschulen zustande kommen können, könnten sich auch unterschiedliche Notenniveaus entwickeln.

Sollte der Erklärungsansatz der unterschiedlichen Eingangseignung für Unterschiede zwischen Studiengängen halten, müssten a) Unterschiede in der durchschnittlichen Abiturnote zwischen den Studiengängen bestehen, die den Unterschieden im Abschlussnotenniveau mindestens tendenziell entsprechen und b) bei Zulassungsbeschränkungen bessere Noten zustande kommen, als in Studiengängen ohne NC, da in diesen Studiengängen die Studienanfänger*innen mit der besten Eingangseignung (gemäß dem stärksten Indikator für diese, der Abiturnote) ‚abgeschöpft' werden und die weniger Begabten, sollten sie nicht Wartesemester sammeln, sich auf andere Studiengänge verteilen.

Tatsächlich ergibt eine einfaktorielle ANOVA hochsignifikante Unterschiede in der Abiturnote für den Faktor Fach/Studiengang. Der Games-Howell Test offenbart signifikante Unterschiede (mindestens auf 5% Niveau) in 52 der 66 Paarvergleiche (Kruskal-Wallis: 50 von 66 Paarvergleichen). Bei einem ersten Blick auf die deskriptive Darstellung ergeben sich zwar zum Teil deutliche Abweichungen in der Rangfolge zwischen Abiturnote und Abschlussnote (Mathematik Lehramt, Biologie, Jura, Maschinenbau), werden jedoch nur die Paarvergleiche mit signifikantem Unterschied jeweils dahingehend betrachtet, ob in diesem Fach/Studiengang eine bessere (schlechtere) durchschnittliche Abiturnote auch mit einer besseren (schlechteren) Abschlussnote gegenüber dem jeweiligen Vergleichsfach/Studiengang einhergeht, so gilt dies für immerhin 38 der 52 Vergleichspaare. Die 14 Paarvergleiche, in denen die bessere (schlechtere) Abiturnote mit der schlechteren (besseren) Abschlussnote einhergeht, sind dabei maßgeblich auf den im Vergleich zur Abschlussnote guten Rangplatz der Abiturnote in Jura sowie auf den vergleichsweise schlechten Rangplatz in Maschinenbau zurückzufüh-

ren. Diese beiden Fächer/Studiengänge sind alleine für 5 der 14 Vergleiche mit nicht passendem Verhältnis zwischen Abitur- und Abschlussnote verantwortlich. Eine studiengangspezifische Selektion nach unterschiedlicher Eingangseignung steht also in Einklang mit den Daten.

Tabelle 69: Durchschnittliche Abiturnote und durchschnittliche Abschlussnote

Studiengang/Fach	Durchschnittliche Abiturnote[a] (Rang)	Durchschnittliche Abschlussnote[b] (Rang)	Übereinstimmende Rangfolge[c]
Mathematik	2.03 (1)	1.59 (3)	9/10
Psychologie	2.07 (2)	1.45 (2)	7/8
Mathematik Lehramt	2.13 (3)	2.04 (8)	5/8
Chemie	2.14 (4)	1.60 (4)	8/9
Biologie	2.22 (5)	1.39 (1)	7/10
Jura	2.30 (6)	3.32 (12)	5/10
Deutsch Lehramt	2.36 (7)	2.11 (9)	7/8
Germanistik Magister	2.37 (8)	1.88 (5)	6/9
Soziologie	2.45 (9)	1.91 (7)	4/6
VWL	2.46 (10)	2.36 (10)	7/8
BWL	2.47 (11)	2.64 (11)	7/9
Maschinenbau	2.51 (12)	1.88 (5)	4/9

[a] Quelle: Konstanzer Studierendensurvey 1983-2010, eigene Berechnungen. Soziologie inkl. Sozialwissenschaften/Sozialkunde; Maschinenbau inkl. Produktions- und Verfahrenstechnik; Chemie inkl. Bio-/Lebensmittelchemie; Mathematik inkl. Statistik.
[b] Stichprobendaten, Durchschnittswerte nur aus den Jahren gebildet, in denen Angaben zum Abitur im Studierendensurvey vorliegen.
[c] nur Paarvergleiche mit signifikant unterschiedlicher Abiturnote berücksichtigt

Werden die Durchschnittsnoten für die (jeweils um bis zu eine Studiendauer gelagten) Jahre, in denen in einem Studiengang an allen Hochschulen des samples einheitlich Zulassungsbeschränkungen existierten oder nicht[93] studiengangübergreifend auf Mittelwertunterschiede getestet, zeigt sich, dass die Noten in zulassungsfreien Studiengängen (n=160) um ca. eine Viertelnote schlechter (zwischen 0.24 (Lag4) und 0.26 (Lag 0) Noten) ausfallen, als in zulassungsbeschränkten (n=119). Dieser stets hochsignifikante Unterschied kommt vermutlich auch deshalb zustande, weil für Psychologie und Biologie, die beiden Studiengänge mit den besten Noten, fast durchgängig hochschulübergreifend Zulassungsbeschränkungen vorlagen, während in den anderen Studiengängen aufgrund uneinheitlicher Regelungen weniger Datenpunkte vorliegen.

Um diese Verzerrung auszuschalten, lässt sich der Anteil Hochschulen mit Zulassungsbeschränkung an allen Hochschulen des jeweiligen Studiengangs im sample berechnen, der den Wert für jede Hochschule in jedem Jahr berücksichtigt und entsprechend der vorliegenden Ausprägungen (0=zulassungsfrei; 1=zulassungsbeschränkt), auf Studiengangebene einen Wert zwischen 0 und 1 annimmt. Hier zeigt sich studiengangübergreifend und unter Kontrolle der Zeitvariablen ein hochsignifikanter negativer Zusammenhang: Je höher der Anteil Hochschulen innerhalb eines Studiengangs mit Zulassungsbeschränkungen, umso besser sind die Abschlussnoten in diesem Studiengang. Die höchste Erklärungskraft weist dabei der um eine ganze Studiendauer gelagte Anteil auf.

[93] Quelle: Hochschulrektorenkonferenz: Studienangebote deutscher Hochschulen (und vorangegangene Serien unter ähnlichen Titeln).

Tabelle 70: OLS-Regression der durchschnittlichen Abschlussnote auf den Anteil zulassungsbeschränkter Hochschulen

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Anteil zulassungsbeschränkte Hochschulen_Lag5	-0.213	0.045	-4.72	0.000
Jahr	-0.010	0.002	-5.20	0.000
Konstante	21.554	3.755	5.74	0.000
				n=392; r^2adj=0.11

Für die Erklärung von Hochschulunterschieden im selben Fach bieten verfügbare Daten hingegen keine Hinweise auf die Wirksamkeit unterschiedlicher Eingangseignung, wie ein Vergleich der Abschlussnoten aus den Berichten des Wissenschaftsrats in Kombination mit Angaben zu lokalen Hochschul-NC-Werten der Zentralen Vergabestelle für Studienplätze (ZVS)[94] zeigt .

Die Standorte mit den höchsten NC-Werten von 26 in Biologie in der ZVS Liste aufgeführten Hochschulen sind im Wintersemester 2005/06, eine Regelstudienzeit (Diplom) vor dem Prüfungsahrgang 2010, die Hochschulen Göttingen (2.0), Bochum (2.1) und Osnabrück (2.1), die mit dem niedrigsten Wert Dresden (1.2), Freiburg (1.2), Saarbrücken (1.3), Hohenheim (1.3) und Karlsruhe (1.3). Die 2010 im Diplom erzielten Noten spiegeln diese Differenz im NC allerdings nicht wider: In Bochum beträgt der Durchschnitt $\bar{x}$=1.5, genauso wie in Hohenheim, Dresden und Freiburg, in Göttingen und Osnabrück beträgt er $\bar{x}$=1.6, genauso wie in Saarbrücken und Karlsruhe. In Darmstadt Düsseldorf, Frankfurt/Main, Kaiserslautern, Kassel und Mainz fanden 2005 alle im Studiengang zugelassenen Bewerber*innen einen Platz. An diesen Hochschulen variiert der Notendurchschnitt 2010 zwischen $\bar{x}$=1.4 und $\bar{x}$=1.7, befindet sich also auf ähnlichem Niveau wie an den Hochschulen mit starker Beschränkung. Dieses Bild bestätigend ergibt eine Korrelationsanalyse zwischen den NC-Werten der Hochschulen und den Durchschnittsnoten einen Wert von r=0.015 (p=0.950), der Mittelwertunterschied zwischen den 20 Hochschulen mit und den sechs ohne lokalen NC beträgt $\bar{x}_1$- $\bar{x}_2$=-0.02 Noten zugunsten der ersten, ist aber nicht signifikant (p=0.614).

In Psychologie war die Zulassung 2005/06 am stärksten in Freiburg (1.0), Heidelberg (1.1) und an der HU Berlin (1.1) beschränkt, in Greifswald, Potsdam und an der FU Berlin lag der NC mit 1.4 am höchsten und in Bielefeld, Braunschweig, Bremen, Chemnitz, Dresden, Darmstadt, Gießen, Koblenz, Magdeburg, Mainz, Marburg, Osnabrück, Saarbrücken und Trier gab es keine örtlichen Beschränkungen für die zugelassenen Bewerber*innen. Die Noten an den stark beschränkten Hochschulen ergeben 2010 in Freiburg einen Durchschnitt von $\bar{x}$=1.3, in Heidelberg von $\bar{x}$=1.4 und an der HU Berlin von $\bar{x}$=1.8. In Greifswald ($\bar{x}$=1.7), Potsdam ($\bar{x}$=1.5) und an der FU Berlin ($\bar{x}$=1.6) sind die Noten etwas schlechter als in Freiburg und Heidelberg, aber etwas besser als an der HU Berlin. Die 40 Standorte mit Informationen zu beiden Werten weisen einen Zusammenhangswert von r=0.377 auf, der auf 10%-Niveau signifikant ist (p=0.058). An den 14 Standorten ohne lokale Beschränkung liegen die No-

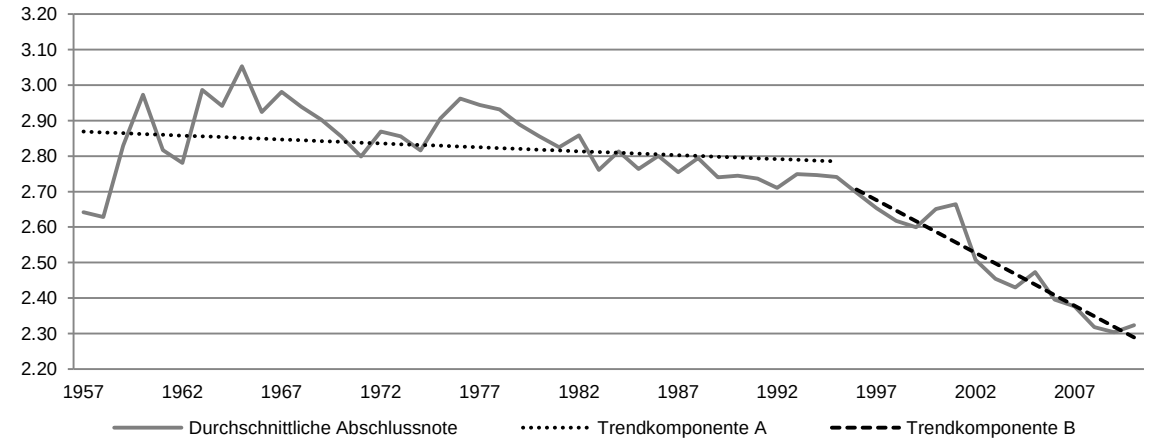
[94] Quelle: ZVS Informations- und Pressestelle (2005): Auswahl- und Verteilungsgrenzen in bundesweit zulassungsbeschränkten Studiengängen zum Wintersemester 2005/2006.

ten im Mittel bei $\bar{x}$=1.55 und damit sogar minimal, wenn auch nicht signifikant (p=0.607) niedriger als an den 26 lokal zulassungsbeschränkten Standorten ($\bar{x}$=1.57).
Der Zusammenhang zwischen NC-Wert und Durchschnittsnote ist im Staatsexamen Medizin ähnlich hoch wie im Diplom Psychologie (r=0.345) und ebenfalls nicht signifikant (p=0.136). Dort weisen 2005/2006 die Hochschulen in Bonn, Freiburg, Heidelberg, Münster und Tübingen mit 1.0 den niedrigsten Numerus Clausus auf, Göttingen, Magdeburg und die Medizinische Hochschule Hannover sind die Standorte mit der höchsten Beschränkungsgrenze (alle 1.4), in Bochum, Duisburg-Essen, Düsseldorf, Frankfurt/Main, Gießen, Halle-Wittenberg, Hamburg, Kiel, Mainz, Marburg, Rostock, Saarbrücken und Ulm gab es 2005 keine lokale Obergrenze. An den stark beschränkten Hochschulen liegen die Noten 2010 zwischen $\bar{x}$=2.2 und $\bar{x}$=2.4. An den weniger stark beschränkten Standorten weisen Göttingen ($\bar{x}$=2.4) und die Medizinische Hochschule Hannover ($\bar{x}$=2.3) ähnliche Ergebnisse auf, nur die Prüflinge in Magdeburg schneiden mit $\bar{x}$=2.6 etwas schlechter ab. Auch an den Hochschulen ohne örtliche Beschränkung liegt der Notendurchschnitt 2010 bei $\bar{x}$=2.3 oder $\bar{x}$=2.4, mit Ausnahme von Kiel ($\bar{x}$=2.6) Halle und Rostock (beide $\bar{x}$=2.5). Die Mittelwertdifferenz zwischen den lokal beschränkten und nicht beschränkten Standorten ($\bar{x}_1$- $\bar{x}_2$=-0.02) ist auch hier nicht signifikant (p=0.741).

Im Zeitverlauf lässt sich der Einfluss der Eingangseignung über den Indikator Zulassungsbeschränkung ebenfalls überprüfen. Sollten Zulassungsbeschränkungen das Leistungs- und damit auch das Notenniveau beeinflussen, müsste mit der Einführung (Aufhebung) einer solchen Beschränkung eine zeitverzögerte Verbesserung (Verschlechterung) der Noten einhergehen. Ein solcher Effekt lässt sich mithilfe einer Interventionsanalyse ermitteln, bei der die Intervention als Dummy-Variable in ein Regressionsmodell der Zeitreihe aufgenommen wird. Interventionen können verschiedene Wirkungsmuster aufweisen. Die Einführung bzw. Abschaffung von Zulassungsbeschränkungen stellt einen sogenannten Stufen-Input dar, bei dem die Intervention dauerhaft wirksam bleibt und eine permanente Niveauverschiebung in der betrachteten Zeitreihe nach sich zieht. Sollte die Einführung (Abschaffung) von Zulassungsbeschränkungen einen Einfluss auf das Notenniveau besitzen, müsste bei einer einmaligen Einführung von Zulassungsbeschränkungen eine (um ca. eine Studiendauer verzögerte) einmalige dauerhafte Niveausenkung der Noten folgen.
Um den Einfluss einer Intervention zu überprüfen, wird eine neue Zeitreihe generiert: Im Falle eines Stufen-Inputs erhalten Messzeitpunkte vor der Intervention dabei die Kodierung Null, Messzeitpunkte ab dem Interventionszeitpunkt die Kodierung Eins. Die neue Zeitreihe muss durchgängig kodiert sein und beide Ausprägungen aufweisen. Dies ist auf Studiengangebene deswegen ein Problem, weil eine Dummy-Kodierung nur für die Jahre möglich ist, in denen an allen Hochschulen des samples einheitlich Zulassungsbeschränkungen vorlagen oder nicht vorlagen.

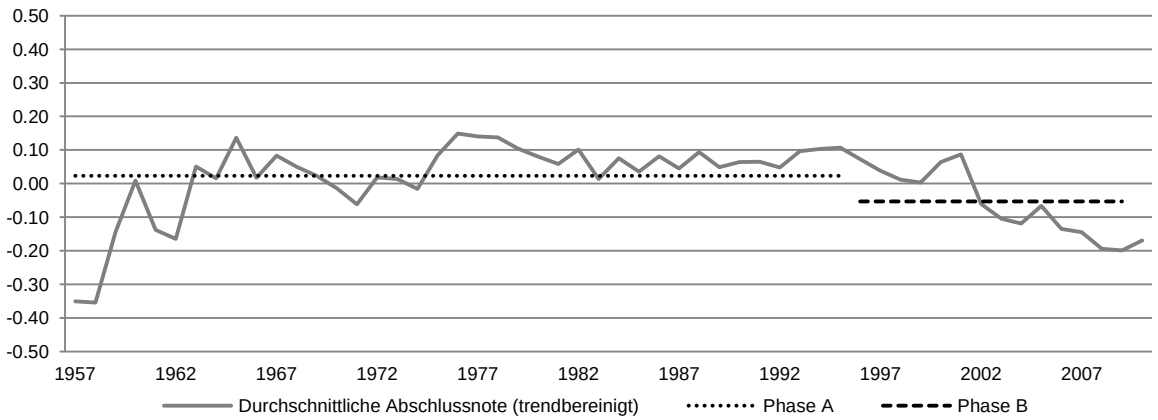
In Verbindung mit dem Erfordernis, dass auch tatsächlich eine Intervention stattgefunden hat und ausreichend Messzeitpunkte davor und danach nutzbar sind, sind auf Studiengangebene nur in BWL alle Voraussetzungen für eine derartige Analyse gegeben. Dort wurden 1991 an allen Hochschulen des samples Zulassungsbeschränkungen eingeführt, die bis 2004, dem letzten Zeitpunkt, für den diese Information vorliegt, Gültigkeit hatten. Da die Studienanfänger*innen von 1991 ca. fünf Jahre später ihren Abschluss gemacht haben, sollte sich ein Interventionseffekt ab 1996 zeigen.

Abbildung 130: Abschlussnoten in BWL und Trendkomponenten vor (A) und ab (B) dem vermuteten Wirkungseintritt



Wie in Abbildung 130 ersichtlich, ist der Trend zu besseren Noten (hier aufgrund des prägnanten Charakters einer linearen Entwicklung als Regressionsgerade einer OLS-Regression der Noten auf die Zeit berechnet) in der um fünf Jahre gelagten Post-Interventionsphase deutlich stärker als in der Prä-Interventionsphase (siehe auch Abschnitt 8.1.2). Da zeitreihenanalytische Verfahren in der Regel stationäre Zeitreihen voraussetzen und der Trend selbst nicht im Fokus der Interventionsanalyse steht, wird in der weiteren Analyse die trendbereinigte Reihe der Noten verwendet, die, wie der Dickey-Fuller Test bestätigt, im Gegensatz zur Originalreihe das Kriterium der Stationarität erfüllt. Die grafische Darstellung der Mittelwerte der trendbereinigten Reihe in der Prä-Interventionsphase und der Phase ab dem um eine Studiendauer gelagten Zeitpunkt der Intervention legt einen Einfluss auf das Notenniveau nahe. Tatsächlich liegt es im Zeitraum 1996 bis 2009 sichtbar niedriger als zuvor.

Abbildung 131: Trendbereinigte Abschlussnoten in BWL und durchschnittliches Niveau der Zeitreihe vor (Phase A) und ab (Phase B) dem vermuteten Wirkungseintritt



Um nun eine mögliche Wirkung der Intervention statistisch erfassen zu können, muss als erstes die zeitliche Struktur der Zeitreihe vor dem (gelagten) Interventionszeitpunkt bestimmt werden. Dies geschieht anhand der Autokorrelationsfunktion (AKF) und der partiellen Autokorrelationsfunktion (PAKF) der Reihe. Die PAKF weist nur einen einzigen signifikanten Koeffizienten, die AKF ein abfallendes Muster auf (siehe Anhang: Abb.A29-A30), weshalb von einem autoregressiven Prozess erster Ordnung (AR (1) auszugehen ist. Dies bedeutet, dass sich der Wert der Zeitreihe zum Zeitpunkt t am besten mit Hilfe des Wertes zum Zeitpunkt t-1 voraussagen lässt. Autoregressive Modelle lassen sich mit dem üblichen OLS-Verfahren schätzen. Die Schätzung der Modelle und die anschließende Diagnose der Modellgüte geben Aufschluss darüber, ob die Struktur der Zeitreihe korrekt erfasst wurde oder ob ein möglicherweise ebenfalls in Frage kommendes anderes Modell die zugrunde liegende zeitliche Struktur besser abbildet. Im Falle der BWL Noten zeigt sich, dass das AR(1) Modell die angemessene Wahl ist. Das Ergebnis der Schätzung lautet:

$$BWL_t = 0.020 + 0.637\ BWL_{t-1} + a_t$$

Das Modell weist ein r^2 von 0.56 auf, was bedeutet, dass alleine mit der Kenntnis des dem Messzeitpunkt vorausgehenden Werts t-1 über die Hälfte der Varianz der Zeitreihe erklärt werden kann. Die Diagnose der Modellgüte des AR(1) Modells zeigt alle zu prüfenden Kriterien erfüllt: Der Lag1 Koeffizient ist hochsignifikant, die Residuen des Modells sind annähernd normalverteilt sowie unkorreliert (und stellen damit einen white-noise Prozess dar) und der Eigenwert der Matrix liegt innerhalb des Einheitskreises.

Die AR(1) Struktur der Zeitreihe der BWL Abschlussnoten gegeben, lassen sich nun die autoregressive Komponente L1 sowie der um eine Studiendauer gelagte Stufeninput als unabhängige Variablen in ein Regressionsmodell aufnehmen. Und es wird deutlich, dass die Intervention tatsächlich einen signifikanten negativen Einfluss auf das Notenniveau besitzt: Eine Studiendauer nach der Einführung der hochschulübergreifenden Zulassungsbeschränkungen in BWL liegen die Noten dort auch unter Kontrolle der autoregressiven Struktur der Reihe und unabhängig vom zeitlichen Abwärtstrend signifikant niedriger als im vorherigen Zeitraum.

Dass die Noten nach der hochschulübergreifenden Einführung der Zulassungsbeschränkungen weiter besser werden, also nicht nur eine Niveauverschiebung sondern eine kontinuierliche Verbesserung zu beobachten ist, steht nicht im Widerspruch zur Annahme einer stufenförmigen Auswirkung der Intervention. Vielmehr ist davon auszugehen, dass sich die konkreten Bedingungen der Zulassung (also zum größten Teil die Abiturnote) nach der Einführung der Zulassungsbeschränkung (die ja aufgrund einer zu hohen Nachfrage gegenüber dem Angebot an Studienplätzen zustande kommt) erst einmal von Jahr zu Jahr verschärfen, bis die Nachfrage nach dem Studiengang wieder sinkt. Entsprechend steht eine fortschreitende Verbesserung der Noten auch nach der Einführung von Zulassungsbeschränkungen in Einklang mit dem aufgezeigten Wirkungsmuster.

Tabelle 71: OLS-Regression der trendbereinigten BWL Noten auf die AR(1) Komponente und den (gelagten) Stufeninput

AV: BWL_trendbereinigt	Koeffizient	Standardfehler	t-Statistik	P>|t|
BWL_trendbereinigt_L1	0.685	0.080	8.57	0.000
Stufeninput_L5	-0.050	0.020	-2.51	0.015
Konstante	0.019	0.010	1.81	0.076
				n=52; r^2adj=0.64

Auch auf Hochschulebene scheitern die meisten Analyseversuche an der Datenstruktur, da die Zeitreihen der Noten entweder einen zu kurzen Prä-Interventionszeitraum beinhalten, um ihre zeitliche Struktur vor dem Interventionszeitpunkt zu ermitteln, nicht genug Messzeitpunkte mit beiden Dummy-Ausprägungen vorliegen oder zu wenige Jahre in der Mitte der Reihe Zulassungsbeschränkungen aufweisen. Lediglich die VWL-Noten in Göttingen erfüllen alle für eine Interventionsanalyse notwendigen Voraussetzungen.

Die PAKF des Prä-Interventionszeitraums weist auch hier nur einen einzigen signifikanten Koeffizienten auf, die AKF kann als abfallend gedeutet werden, wobei aufgrund der Kürze der Reihe kein so eindeutiges Muster zu erkennen ist, wie bei den BWL Noten (siehe Anhang: Abb.A31-A32). Das Korrelogramm lässt jedoch keinen anderen Prozess erkennen und die Modelldiagnose des AR(1) Modells mit der Schätzung

$$VWL\_Gö_t = 0.017 + 0.495\ VWL\_Gö_{t-1} + a_t$$

bestätigt auch hier die Wahl eines AR(1) Modells.

Das Modell weist ein r^2 von 0.25 auf, was bedeutet, dass alleine mit der Kenntnis des dem Messzeitpunkt vorausgehenden Werts t-1 ein Viertel der Varianz der Zeitreihe erklärt werden kann. Werden die autoregressive Komponente L1 sowie der um eine Studiendauer gelagte Stufeninput als unabhängige Variablen in ein Regressionsmodell aufgenommen, bringt die Intervention zwar das erwartete negative Vorzeichen hervor, der Effekt ist aber nicht signifikant.

Tabelle 72: OLS-Regression der trendbereinigten VWL Noten auf die AR(1) Komponente und den (gelagten) Stufeninput

AV: VWL_Gö_trendbereinigt	Koeffizient	Standardfehler	t-Statistik	P>|t|
VWL_Gö_trendbereinigt_L1	0.481	0.129	3.73	0.001
Stufeninput_L5	-0.053	0.047	-1.12	0.270
Konstante	0.017	0.025	0.69	0.494
				n=46; r^2adj=0.25

Zwei Beispiele, in denen nicht ausreichend Messzeitpunkte vor dem Interventionszeitpunkt vorhanden sind, lassen sich im Studiengang Chemie finden. Hier wurden in Berlin (1970) und in Göttingen (1971) Zulassungsbeschränkungen eingeführt, die allerdings an beiden Hochschulen nur wenige Jahre später (1976) wieder aufgehoben wurden. Zwar sollten fünf bzw. sechs Jahre, in denen die Zulassung begrenzt war, ausreichen, um im Zeitraum nach der Intervention (in diesem Fall die Aufhebung der Beschränkungen) einen nachvollziehbaren Effekt zu produzieren, allerdings reicht der Zeitraum bei weitem nicht aus, um eine möglicherweise zugrundeliegende Prozessstruktur zu identifizieren. Der grafische Vergleich des (um fünf Jahre gelagten) Notenniveaus beider Zeiträume legt jedoch nahe,

dass auch die Abschaffung von Zulassungsbeschränkungen nicht ohne Wirkung auf die Notenhöhe bleibt, die Noten dann schlechter werden:

Abbildung 132: Verlauf der Abschlussnoten in Chemie an der FU Berlin (durchgehende Linie) und durchschnittliches Niveau der Zeitreihe vor (Phase A) und ab (Phase B) dem vermuteten Wirkungseintritt

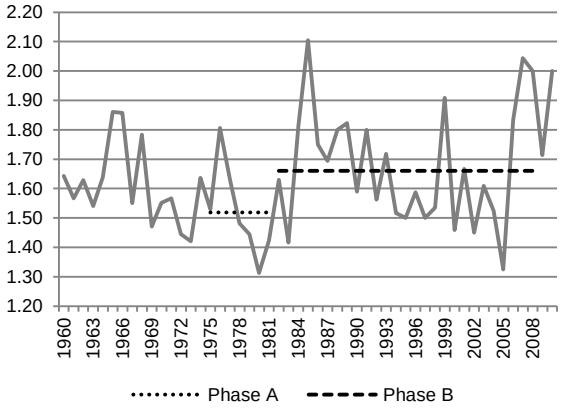
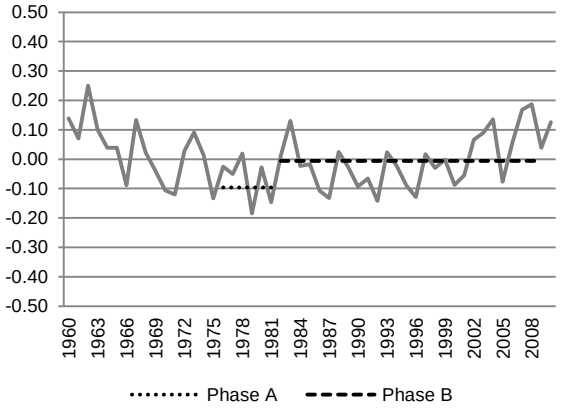


Abbildung 133: Verlauf der trendbereinigten Abschlussnoten in Chemie an der Uni Göttingen (durchgehende Linie) und durchschnittliches Niveau der Zeitreihe vor (Phase A) und ab (Phase B) dem vermuteten Wirkungseintritt



Auch wenn der Stufeninput im Interventionsmodell der Göttinger VWL-Noten nicht signifikant ist, deuten die Daten zumindest darauf hin, dass die Eingangseignung im Längsschnitt einen Zusammenhang mit dem Notenniveau aufweist. Mindestens stehen die Resultate im Einklang mit dieser Annahme, die ja auch schon durch den bekannten Zusammenhang zwischen Abiturnote und Studienerfolg gestützt wird.

Selbstselektion im Zeitverlauf

Neben der Steuerung von Studierendenzahlen (und damit der Steuerung von Leistungsniveaus) über Zulassungsbeschränkungen als nachfrageinduziertem Fremdselektionsmechanismus ist ebenfalls zu bedenken, dass das Notenniveau im Zeitverlauf theoretisch auch durch die Verteilung der Studierenden auf Studiengänge, Hochschulen oder Kurse/Teilprüfungen mit stabilen Unterschieden im Notenniveau beeinflusst werden kann. Zum Beispiel könnten in einem beliebigen Studiengang an einigen Hochschulen in der Regel deutlich bessere Noten als an den übrigen Standorten vergeben werden. Entwickelt sich unter Studienanfänger*innen durch eine zunehmende Informationsdichte (z.B. durch Informationsplattformen und Forendiskussionen im Internet) ein Bewusstsein für eine solche Polarisierung im Notenniveau, könnte dies bei einer ebenfalls im Zeitverlauf zunehmenden regionalen Mobilität von Studienanfänger*innen (Kultusministerkonferenz 2011) und einer zeitgleichen Priorisierungstendenz, die den Erwerb von Bildungszertifikaten über den Bildungserwerb stellt (vgl. Kirp 2003; Rojstaczer/Healy 2012; Rosovsky/Hartley 2002), einen Anreiz für Studierende bieten, zunehmend die Hochschulen mit besseren Noten bei der Standortwahl zu bevorzugen. Solange die Nachfrage das Angebot in diesem Fall nicht übersteigt und keine Reglementierung durch überregionale Zulassungsbeschränkungen erfolgt, würde eine solche Selbstselektion in Hochschulen mit besseren

Noten den Notendurchschnitt im Studiengang senken, ohne dass das individuelle Verhältnis von Leistung und Bewertung sich ändert.
Hinweise auf derartige Prozesse können die relativen Studierendenzahlen bieten: Sollte etwa eine im Zeitverlauf zunehmende Selbstselektion in Studiengänge mit besseren Noten existieren (was aufgrund fachspezifischer Motivationsmuster bei Studienbeginn (siehe Abschnitt 3.2.2) nicht anzunehmen ist), müssten die Studienanfänger*innenzahlen in Studiengängen mit guten Notenniveau im Verhältnis zu denen mit vergleichsweise schlechtem Niveau zunehmen.
Eine derart produzierte Notenverbesserung im Zeitverlauf könnte aber sowieso nur für über mehrere Studiengänge oder Fächer aggregierte Notenlevel erklärungsrelevant sein. Naheliegender ist es, dass Noten bei der Hochschulwahl und vor allem innerhalb einer Universität bei der Wahl von Teilprüfungen einen Anreiz bzw. eine Hemmung darstellen. Auch hier könnte die Entwicklung der relativen Studierenden- bzw. Prüfungszahlen (innerhalb eines Studiengangs bzw. innerhalb eines Studiengangs an einer Hochschule) einen Hinweis darauf geben, ob eine Selbstselektion nach Noten existiert.

Auf Studiengangebene stehen unter dem Label „Große Hochschulstatistik“ jährlich verschiedene Fachserien hochschulstatistischer Daten des Statistischen Bundesamtes, unter anderem zu den Studierendenzahlen nach erstem Studienfach, zur Verfügung, anhand derer eine Verschiebung von Studienanfänger*innenanteilen überprüfbar ist. Leider sind die Daten im Zeitverlauf nicht einheitlich erfasst: Neben weniger schwerwiegenden (historisch bedingten) Unterschieden in der Abgrenzung der erfassten Hochschularten und der Studierenden[95] ist vor allem problematisch, dass für den Zeitraum von 1972 bis 1977 angehende Lehrer*innen nicht nach dem ersten Studienfach, sondern nach ihrem Abschluss als Lehramtsstudierende erfasst wurden, davor und danach jedoch auch ihrem ersten Studienfach zugerechnet wurden. Ab 1978 werden Lehramtsstudierende zwar noch einmal zusätzlich nach erstem Studienfach aufgeführt, allerdings nur über alle Hochschularten gemittelt. Dies hat zur Folge, dass a) keine differenzierten Informationen für die erfassten Studiengänge Mathematik und Germanistik auf *Lehramt an Gymnasien* ausgelesen werden können und b) die Zahlen der Diplom- bzw. Magisterstudierenden über alle Hochschulen zusammengefasst verwendet werden müssten, um sie vom Anteil Lehramtsstudierende bereinigen zu können. Es existieren jedoch nicht nur deutliche Unterschiede in den Studierendenzahlen zwischen Hochschulen und Fachhochschulen (und

[95] Neben der Darstellung der Studierendenzahlen über alle Hochschularten gemittelt sind differenzierte Angaben enthalten, die es ermöglichen, nur wissenschaftliche Hochschulen (bis 1970) bzw. wissenschaftliche Hochschulen und Kunsthochschulen (1971-1974) bzw. Universitäten und Technische Hochschulen (ab 1975) auszulesen. Vor 1971 beschränken sich die nach erstem Studienfach gegliederten Informationen zudem auf Studierende mit deutscher Staatsbürgerschaft. Die Werte für das Wintersemester 1973/74 liegen nicht vor und wurden linear interpoliert, da dieser Jahrgang der Fachserie nicht erschienen ist (vergleiche mit den Beständen der Zeitschriftendatenbank (ZDB)).

auch in den dort vergebenen Noten – siehe Grözinger 2015), eine entsprechende Datenbereinigung wäre auch nur für den Zeitraum ab 1978 überhaupt möglich - nicht jedoch für die Zeit vor 1972.

Daher ist es sinnvoller, statt der über alle Hochschultypen gemittelten Daten für alle Fächer im sample nur die Angaben zu den wissenschaftlichen Hochschulen zu verwenden und sich dafür bei der Auswertung auf die Fächer ohne relevanten Lehramtsanteil (Jura, BWL, VWL, Psychologie, Soziologie, Maschinenbau) zu beschränken. Ließen sich in den Studiengängen mit vergleichsweise guten (schlechten) Noten im Zeitverlauf steigende (sinkende) Anteile an Erstsemestern nachweisen, würde dies mit der Selbstselektionsthese in Einklang stehen.

Wird für diese Fächer (abschlussübergreifend) die Abstufung in der Höhe der Anteile mit der Abstufung nach Notenniveau verglichen, fällt ins Auge, dass die Fächer mit den höchsten Anteilen, Jura und BWL, auch die schlechtesten Notenniveaus aufweisen, während in Psychologie geringe Erstsemesteranteile ebenso die Regel sind, wie sehr gute Noten. Ein relativer Anstieg der Studienanfänger*innen in Psychologie aufgrund der stets sehr guten Noten ist aber nicht zu erwarten: Durch das Instrument der Studienplatzbegrenzung und die in Psychologie schon immer strengen Zulassungsbeschränkungen wird die Möglichkeit eines solchen Effekts hier bereits in der Theorie ausgehebelt. In Hinblick auf die übrigen Fächer muss zudem a) beachtet werden, dass zwar deren Anteile an den Erstsemestern aller Studiengänge (ohne Lehramt) berechnet werden können, allerdings weder Informationen über die Entwicklung der Anteile der nicht im sample enthaltenen sowie der bei den Analysen nicht berücksichtigten Studiengänge, noch der dortigen Noten vorliegen[96] und b) dass durch die Einführung neuer (nicht erfasster) Studiengänge im Zeitverlauf in allen Studiengängen unabhängig von anderen Einflussfaktoren mit einer Verringerung der Anteile im Zeitverlauf zu rechnen ist[97]. Verschiebungen zwischen den Studiengängen müssten aber auf Differenzen im Notenniveau zurückgeführt werden und können nicht am absoluten Notenniveau abgelesen werden.

Unabhängig von diesen Einschränkungen sollte bei Zutreffen der Selbstselektionsthese ein Absinken der Erstsemesteranteile in Jura, BWL und VWL zu beobachten sein - diese Studiengänge gehören auch in den vorliegenden Vergleichsstudien immer zu denen mit den schlechtesten Notenniveaus (siehe Abschnitt 6.2) und sollten Studierende tendenziell eher abschrecken, sollte es zutreffen, dass sie ihre Fachwahl in zunehmendem Ausmaß von der zurückliegenden Notengebung abhängig machen. Stattdessen zeigen sich in Jura und BWL starke zyklische Bewegungen in den Anteilen an Studienanfänger*innen ohne Trend nach oben oder unten. Ließen sich die neu entstehenden Studiengänge, die sich im Zeitverlauf auf Kosten der Erstsemesteranteile der klassischen Studiengänge etablieren, einberechnen, wäre vermutlich sogar ein relativer Nettozuwachs zu verzeichnen. In VWL ist da-

[96] Der Anteil der Erstsemester in den sechs hier berücksichtigten Fächern an der Gesamtheit aller Erstsemester (ohne Lehramt) beträgt aufsummiert über alle Jahre gemittelt 22.3%.

[97] Der Anteil der Erstsemester in den sechs hier berücksichtigten Fächern an der Gesamtheit aller Erstsemester beträgt im Jahr 1959 noch 28.3%, 2010 liegt er nur noch bei 19.4%.

gegen neben ebenfalls vorhandenen, wenn auch schwächeren Zyklen ein leichter Abwärtstrend zu erkennen (Abb.134). Wird die Stärke dieses Trends In Relation zur Stärke des Abwärtstrends in den Anteilen aller sechs berücksichtigten Studiengänge gesetzt, zeigt sich jedoch, dass das Absinken der VWL-Anteile nicht stärker ausfällt als der generelle Rückgang der summierten sechs Anteile an der Gesamtheit aller Studienanfänger*innen. Dies deutet daraufhin, dass der Rückgang an Erstsemestern in VWL eher der zunehmenden Diversifizierung des Studienangebots geschuldet ist.

Abbildung 134: Fachspezifische Erstsemesteranteile im Zeitverlauf (LOWESS 0.3, nur Universitäten, ohne Lehramt)

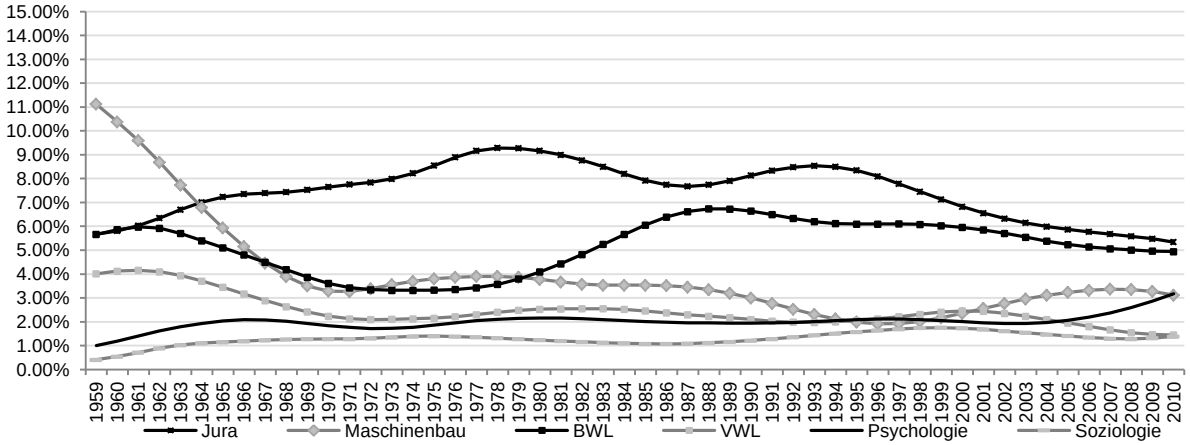


Abbildung 135: Fachspezifische Abschlussnoten im Zeitverlauf (LOWESS 0.3)

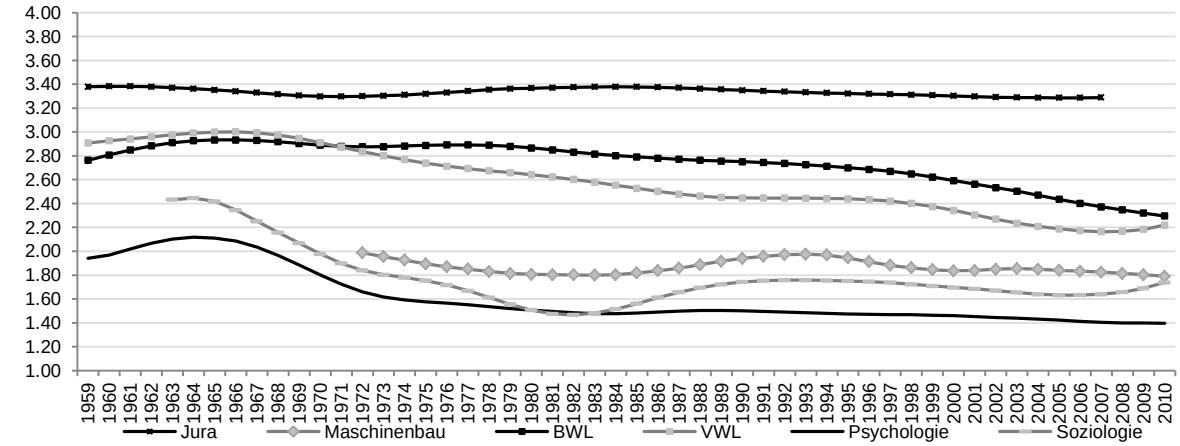
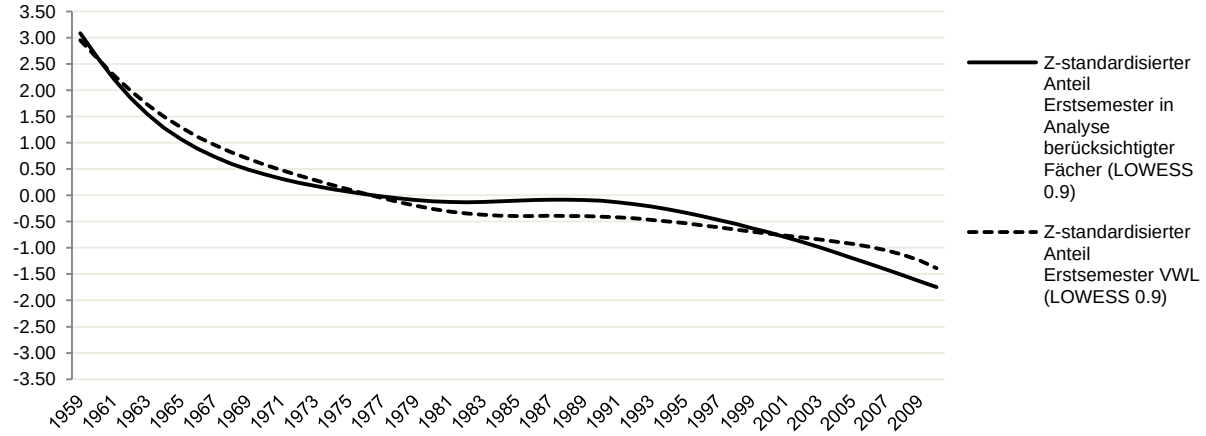


Abbildung 136: Z-Standardisierte Trendkomponenten der Erstsemesteranteile (LOWESS 0.9)



Auf Hochschulebene lässt sich eine zunehmende Verschiebung der Studierendenanteile innerhalb eines Studiengangs prinzipiell mithilfe der um eine Studiendauer in die Vergangenheit verschobenen Prüfungszahlen (welche in etwa den Studienanfänger*innenzahlen der Prüfungskohorte abzüglich Studienabbrecher*innen und Fach- bzw. Hochschulwechsler*innen entsprechen) überprüfen, indem sie in Relation zu den Erstsemesterzahlen der Hochschulstatistik für den ganzen Studiengang gesetzt werden. Dabei wird wieder auf die Analyse der Fächer mit starkem Lehramtsanteil, außerdem der Maschinenbau- (nur zwei Hochschulen im sample), Jura- (Angaben nur auf Bundeslandebene) und Soziologiedaten verzichtet. Letztere sind abschlussübergreifend erfasst und es müsste zur Prüfung der Selbstselektionsthese zunächst geklärt sein, ob die Wahl eines Diplom- oder Magisterstudiengangs Soziologie bei interessierten Studienanfänger*innen möglicherweise die Hochschulwahl beeinflusst (oder umgekehrt). Damit verbleiben nur VWL und BWL für die Analyse.
Auch auf Hochschullevel liegen keine Informationen über die Entwicklung der Anteile der nicht im sample enthaltenen Hochschulen desselben Studiengangs und der dort vergebenen Noten vor[98] und hier ist ebenfalls zu beachten, dass eine wachsende Anzahl an Hochschulen unabhängig von allen anderen Einflüssen zu einem Absinken der Anteile im Zeitverlauf geführt haben dürfte [99]. Infolgedessen kann auch hier nur überprüft werden, ob sich in Hochschulen mit auffällig schlechtem bzw. auffällig gutem Notenniveau Veränderungen der Erstsemesteranteile im Zeitverlauf zeigen.
Für Psychologie erübrigt sich die Analyse an dieser Stelle. Denn bei der starken Begrenzung der Studienplätze und der damit einhergehenden konstant vollständigen Auslastung im Studiengang können hochschulspezifische Veränderungen der Erstsemesteranteile nur aufgrund von Veränderungen in der Anzahl der angebotenen Plätze an den einzelnen Hochschulen zustande kommen und nicht durch selbstinduzierte Wanderbewegungen der angehenden Studierenden, die ja zentral verteilt werden.

In VWL gibt es keine deutlichen Abweichungen mit besonders hohem Notenniveau, die besten Noten werden in der Regel in Berlin und Karlsruhe vergeben. Da der Vergleich mit den Daten von Hitpass und Trosien und dem Wissenschaftsrat für den Zeitraum von 1953 bis 2000 impliziert, dass dies auch im Vergleich zu den hier nicht erfassten Standorten so ist, sollte ein Anstieg der relativen Anzahl an Studienanfänger*innen im Zeitverlauf stattfinden. Unter Berücksichtigung des Umstands, dass die Erweiterung der Hochschullandschaft zu einem generellen Absinken der Anteile an allen Hochschulen der Stichprobe führen dürfte, müsste immer noch ein schwächerer Abwärtstrend als im Durchschnitt

[98] Die Anteile der (über die Prüfungszahlen ermittelten) Erstsemester an den hier berücksichtigten Hochschulen an der Gesamtheit aller Erstsemester im jeweiligen Fach (ohne Lehramt) betragen aufsummiert über alle Jahre gemittelt: 18.3% in Psychologie, 17.0% in VWL, 15.9% in BWL und 12.4% in Soziologie.
[99] Die Anteile der (über die Prüfungszahlen ermittelten) Erstsemester an den hier berücksichtigten Hochschulen an der Gesamtheit aller Erstsemester im jeweiligen Fach (ohne Lehramt) sinken zwischen 1959 und 2005 in Psychologie von 24.5% auf 7.2%, in VWL von 37.3% auf 7.0%, in BWL von 28.5% auf 3.6% und in Soziologie von 29.0% auf 3.0%.

der restlichen Standorte zu beobachten sein. In Berlin sinkt der Anteil der Studierenden jedoch ähnlich stark, in Karlsruhe sogar noch stärker. Für eine zunehmende Attraktivität der Hochschulen aufgrund des vergleichsweise guten Notenniveaus gibt es demnach keine Anzeichen (Abb.137-139).

Auch in BWL gibt es nach oben keine Ausreißer und wieder ist Berlin der vielversprechendste Standort, um Hinweisen auf einen Selbstselektionsprozess auf Hochschullevel nachzugehen: Das Notenniveau liegt deutlich unter dem der anderen Stichprobenstandorte (allerdings aufgrund der vermuteten Überschätzung der BWL-Noten im sample nur zwischen 1963 und 1996 auch geringfügig unter dem der Vergleichsstudien). Bei den betriebswirtschaftlichen Erstsemesterzahlen geht der Anteil der Berliner Studierenden mit der Zeit nicht so stark zurück wie in der Summe der Hochschulen Göttingen und Münster (Abb.140 bis 142, Tübingen ist bei der Addition der Erstsemesteranteile aufgrund der Kürze der Zeitreihe nicht berücksichtigt worden). Dies steht in Einklang mit der Selektionsannahme, allerdings sind die Vergleichsmöglichkeiten mit nur zwei weiteren Hochschulen auch begrenzter als in VWL mit vier Hochschulen.

Abbildung 137: Hochschulspezifische Erstsemesteranteile im Fach VWL im Zeitverlauf (LOWESS 0.3)

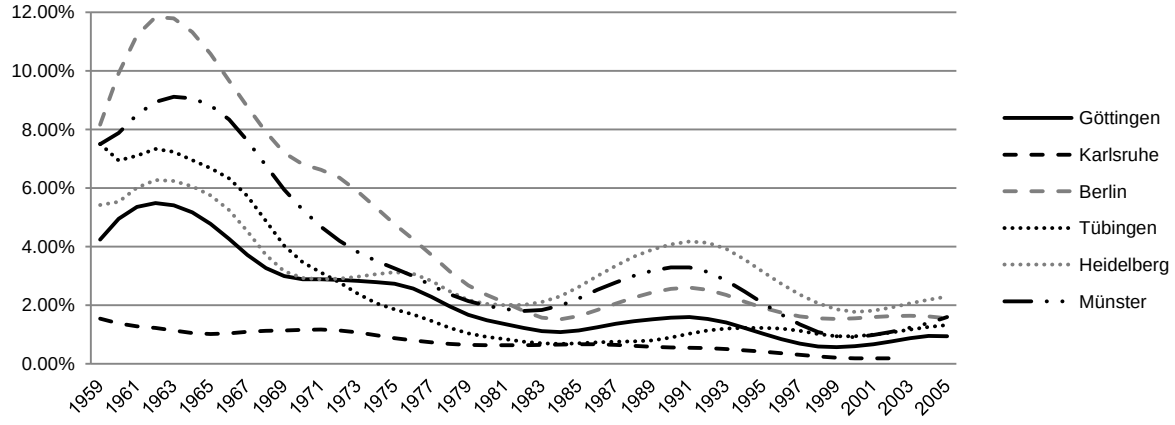


Abbildung 138: Hochschulspezifische Abschlussnoten im Studiengang VWL Diplom im Zeitverlauf (LOWESS 0.3)

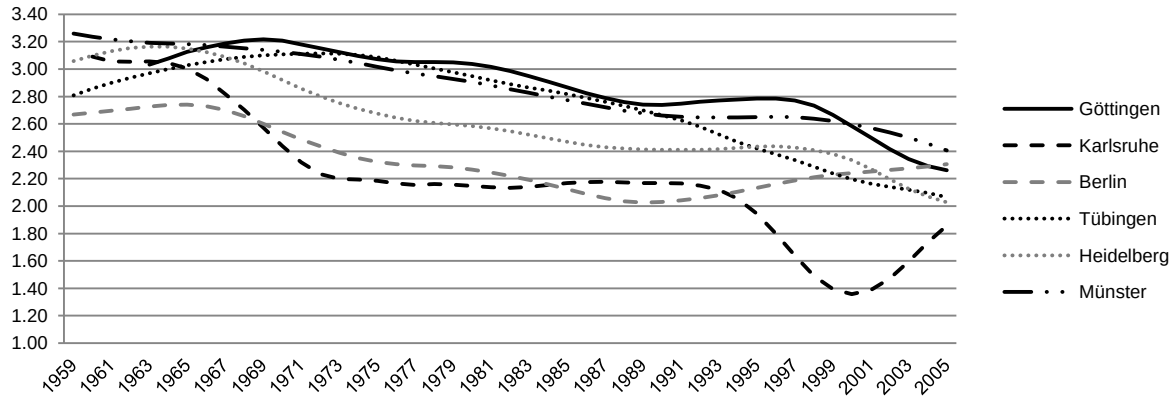


Abbildung 139: Z-Standardisierte Trendkomponenten der Erstsemesteranteile (LOWESS 0.9) im Fach VWL im Zeitverlauf

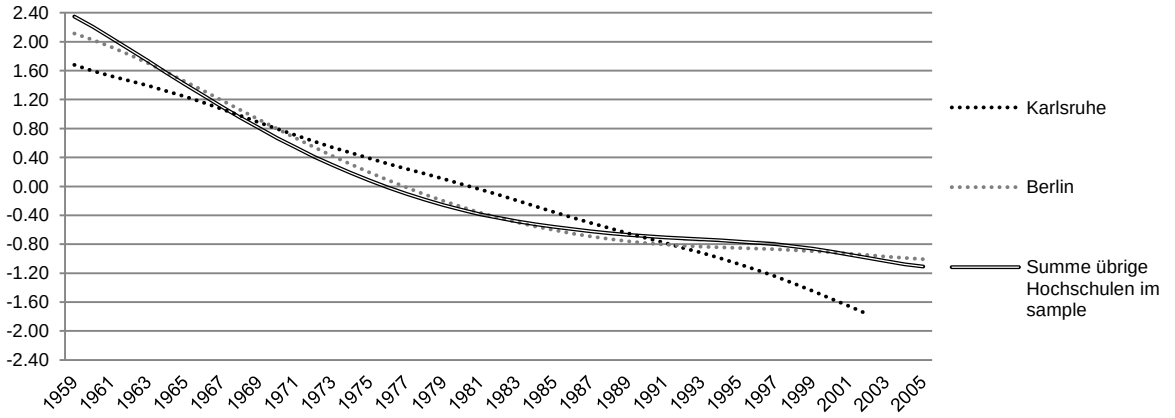


Abbildung 140: Hochschulspezifische Erstsemesteranteile im Fach BWL im Zeitverlauf (LOWESS 0.3)

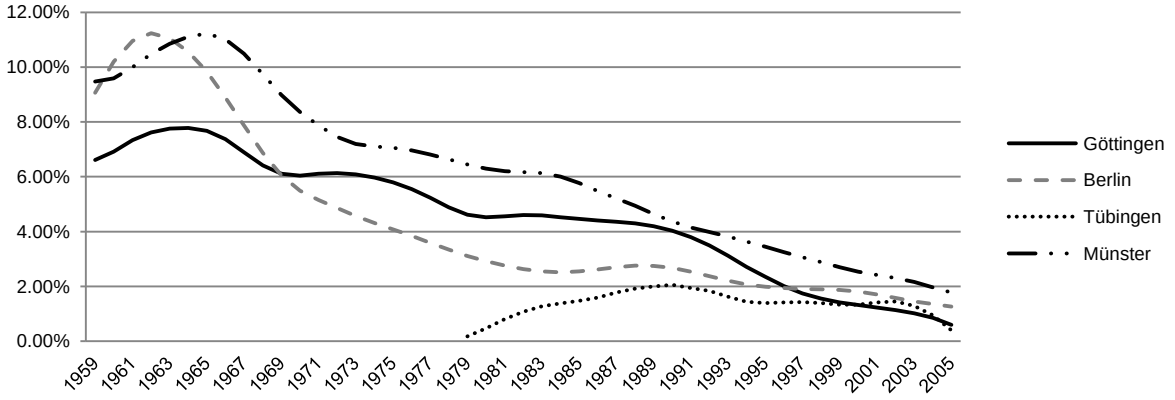


Abbildung 141: Hochschulspezifische Abschlussnoten im Studiengang BWL Diplom im Zeitverlauf (LOWESS 0.3)

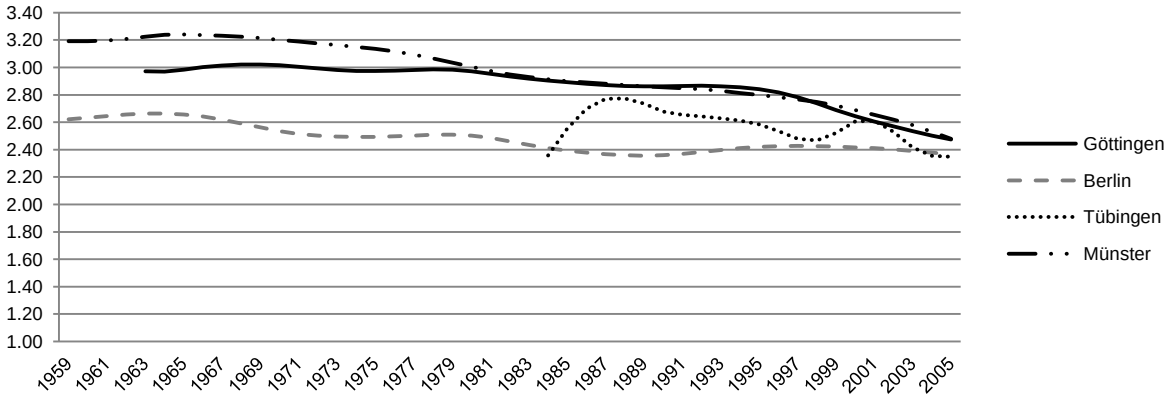
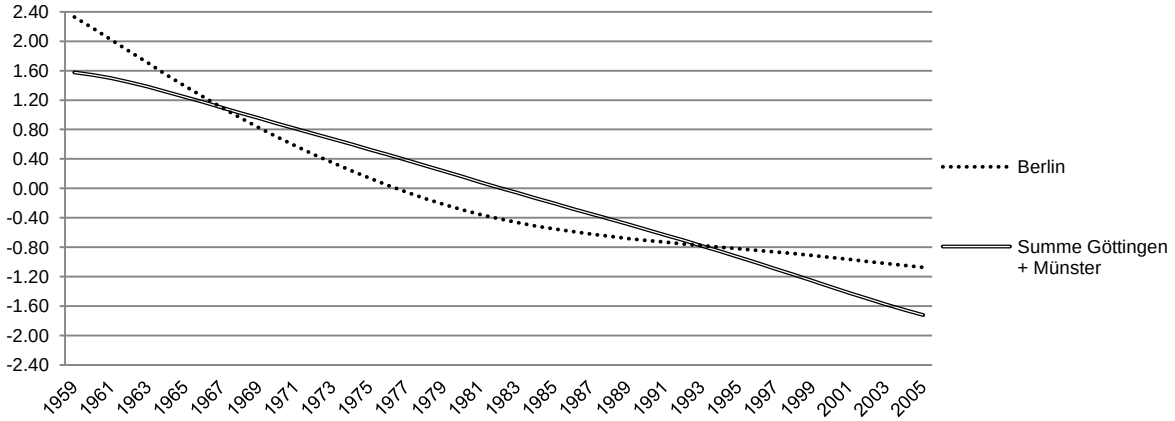


Abbildung 142: Z-Standardisierte Trendkomponenten der Erstsemesteranteile (LOWESS 0.9) im Fach BWL im Zeitverlauf



Die Einschränkungen, denen die Interpretation der verfügbaren Erstsemesterzahlen in Abhängigkeit der Notenentwicklung unterliegt, sind nicht gerade unbedeutend. Dennoch müsste sich ein Selbstselektionsprozess - auch bei einem allgemeinen Absinken der Erstsemesteranteile - mindestens anhand geringerer Trendstärken bei besseren Notenniveaus äußern. Dies ist aber lediglich im Hochschulvergleich im Beispiel BWL der Fall. Belegt ist die These damit auch bei der Hochschulwahl nicht - in Anbetracht der Tatsache, dass sie für das verwandte Fach VWL widerlegt werden kann, ist die Aussagekraft der in Einklang mit der Annahme stehenden BWL-Daten bei geringeren Vergleichsmöglichkeiten als in VWL eher begrenzt.

Ob innerhalb eines Studiengangs an einer Hochschule bestimmte Teilprüfungen, die als Wahl-(pflicht)prüfungen angeboten werden aufgrund ihres besseren Durchschnitts zunehmend häufiger gewählt werden, lässt sich an der Entwicklung der relativen Häufigkeiten ablesen, mit denen Studierende bestimmte Teilprüfungen ablegen, wenn sie eine Auswahl haben. Je nach Studiengang beschränkt sich die Auswahl entweder auf fachinterne oder fachexterne Teilprüfungen, gelegentlich ist es auch möglich, zwischen fachinternen und -externen Prüfungen zu wählen.
Überprüft werden kann eine Verschiebung im Zeitverlauf nur, wenn für die Studiengänge an den einzelnen Hochschulen detaillierte Informationen zu den belegten Prüfungen vorliegen und im Zeitverlauf konstante Wahlmöglichkeiten gegeben sind. Die Änderung von Prüfungsordnungen ist ein beispielhafter Faktor, der die Analyse der zeitlichen Entwicklung auf bestimmte Zeitabschnitte beschränkt. Zudem muss die Auswahl verschiedener Teilprüfungen so gestaltet sein, dass jeweils ausreichende Fallzahlen vorliegen - belegen stets um die 90 von 100 Prüflingen die gleiche Teilprüfung, während sich die restlichen 10 auf fünf verschiedene Alternativen verteilen, sind dies schlechte Voraussetzungen für einen informationsbasierten Selektionsprozess. Grundsätzlich von den Analysen ausgeschlossen wurden die Rechtswissenschaften, zu denen keine detaillierten Prüfungsinformationen vorliegen, Maschinenbau, da hier eine zu große Auswahl an Prüfungsbelegungen existiert, um sie sinnvoll auf verschiedene Wahlmöglichkeiten zu aggregieren sowie die beiden Lehramts- und Magisterstudiengänge, da auch hier nur Informationen zu den belegten Fächern, nicht aber zu den Teilprüfungen vorliegen. Der Zeitraum nach 1997 kann nicht analysiert werden, da in der Hochschulprüfungsstatistik keine Ergebnisse zu Teilprüfungen enthalten sind. Im Gegensatz zur Fach- bzw. Studiengang- und Hochschulwahl stehen vollständige Informationen über alle gewählten Teilprüfungen zur Verfügung und es lassen sich genaue Differenzen im Notenniveau zwischen den Teilprüfungsnoten berechnen.
Grundsätzlich geeignet für die Analyse sind über die restlichen Studiengänge und alle Hochschulen betrachtet 18 Wahl-(pflicht)prüfungen. Von diesen 18 Prüfungen weisen acht keine im Zeitverlauf

andauernden Unterschiede im Notenniveau zwischen den Wahlmöglichkeiten auf, so dass dort kein Anreiz besteht, taktisch zu wählen. Für die übrigen 10 Prüfungsteile, in denen mindestens zwei Wahlmöglichkeiten zeitlich stabile Notenunterschiede aufweisen, ist im Folgenden jeweils sowohl die Entwicklung der dort erzielten Durchschnittsnote als auch die Entwicklung des jeweiligen Anteils an Prüflingen an der Gesamtheit der Prüflinge, die eine dieser Wahlmöglichkeiten belegt hat, grafisch dargestellt (Abb.143-Abb.162). Und die grafische Analyse reicht bereits aus, um zu erkennen, dass unterschiedliche Notenniveaus in verschiedenen Wahlprüfungen keine relevanten Verschiebungen in den Anteilen, zu denen sie jeweils belegt werden, nach sich ziehen. In sechs der 10 Vergleiche von Noten- und Prüflingsanteilsentwicklung bleiben die Anteile der Wahlprüfungen mit besserem (schlechterem) Notenniveau langfristig relativ stabil oder sinken (steigen) sogar noch.

Im Diplomstudiengang Mathematik an der FU Berlin sowie an der Universität Münster (1984-1997) passt die anteilsmäßige Entwicklung der Prüfungszahlen auf den ersten Blick (in Berlin zumindest für einen Abschnitt der Reihe) mit der Annahme taktischen Wahlverhaltens überein. In beiden Fällen steigt der Anteil derjenigen, die Informatik bei einem vergleichsweise niedrigen Notenniveau als Wahlprüfung wählen (in Berlin gegenüber Physik, in Münster gegenüber Physik, VWL und BWL). In Berlin hält dieser Trend jedoch auch noch an, nachdem sich das Notenniveau der beiden Nebenfächer zunächst angleicht und anschließend sogar umkehrt, in Münster tritt der Effekt nur in Informatik auf, nicht aber in Physik, wo ein vergleichbares Notenniveau herrscht. Der langfristige Anstieg des Anteils an Nebenfachprüfungen in Informatik muss wohl eher als Folge der zunehmenden Etablierung der Informatik als Wissenschaft verstanden werden und kann damit auf die Einführung eines neuen Teilprüfungsangebotes, dessen Wahl einer von Noten unabhängigen Verwertungsperspektive geschuldet ist, zurückgeführt werden. In VWL an der FU Berlin geht das in den Politikwissenschaften als Nebenfach im Vergleich zu den Geschichtswissenschaften bessere Notenniveau zunächst mit einem Anstieg des Anteils an Prüfungen einher, bevor dieser Anteil - ebenso wie die Differenz im Notenniveau - ab Anfang der 1990er Jahre sinkt. Diese durchaus mit der These taktischer Prüfungswahl konforme Entwicklung wird allerdings von einem durchgängig besseren Notenniveau in Politik begleitet, welches auch nach der Annäherung der Notenniveaus noch sichtbar unter dem in Geschichte liegt - während die Anteile der Prüfungen sich in Politik nach dem Peak immer weiter verringern. Auch hier verläuft die Entwicklung also nur dem ersten Anschein nach, wie es bei einer Selbstselektion in die vergleichsweise besser benotete Wahlprüfung zu erwarten wäre.

Lediglich in den externen Wahlmöglichkeiten der BWL Prüfungen an der Universität Göttingen findet sich eine passende Entwicklung: Das im Zeitverlauf im Vergleich zu Jura (Privatrecht) und Wirtschafts- und Sozialgeschichte langfristig stabil bessere Notenniveau in der Wirtschafts- und Sozialpsychologie (WiSo-Psychologie) geht einher mit einem stetigen Anteil an Prüflingen. Inwiefern dieses Muster mehr als einen scheinbaren Zusammenhang darstellt, lässt sich statistisch überprüfen: Der Anteil an

Prüflingen müsste sich entsprechend der in den Jahren zuvor bestehenden Differenzen im Notenniveau zwischen den Wahlfächern entwickeln. Mit zunehmender Notendifferenz zwischen zwei Wahlmöglichkeiten sollte sich der Anteil an Prüflingen zeitverzögert zugunsten der Option mit dem besseren Notenniveau verschieben. Ein solch zeitverzögerter Zusammenhang lässt sich auch ohne eine theoretisch plausible Begründung, bei welchem Lag dieser wirksam wird, mit Hilfe von Vektorautoregressionsmodellen (VAR) berechnen.
Die Regressionskoeffizienten der abhängigen Variable werden in VAR-Modellen mit Hilfe der vergangenen Werte aller im Modell enthaltenen Variablen geschätzt. Auf diese Weise lassen sich zeitverzögerte Zusammenhänge zwischen Zeitreihen aufdecken Dabei werden alle Variablen sowohl als abhängige als auch als unabhängige Variablen behandelt, so dass nicht eine vorherige theoretische Herleitung die Richtung des Zusammenhangs vorgibt, sondern die Richtung aus der Datenstruktur ‚ausgelesen' wird (Brandt/Williams 2007).

Der zu überprüfende Zusammenhang zwischen der Notendifferenz zwischen Jura und Wirtschafts- und Sozialpsychologie und dem jeweiligen Anteil an Prüfungen muss in beiden Fällen mit den ersten Differenzen der Reihen berechnet werden, da die Prüfungsanteile einen deutlichen Trend aufweisen und auch durch Trendbereinigung nicht in stationäre Reihen überführt werden können, was Voraussetzung für die Verwendung in VAR-Modellen ist (Thome 2005). Sowohl der Anteil Prüfungen in Jura als auch der Anteil Prüfungen in Wirtschafts- und Sozialpsychologie als abhängige Variable in VAR-Modellen weisen dabei die erwartete Beziehung zur Notendifferenz zwischen den Prüfungsgebieten auf: In beiden Fällen zeigen die stabilen VAR-Modelle signifikante erste Leads - wird der Prüfungsanteil in Jura als abhängige Variable genutzt (Tabelle 73, oben), ist das Vorzeichen des Lead1-Koeffizienten der Notendifferenz negativ, wird der Prüfungsanteil in Wirtschafts- und Sozialpsychologie auf die Notendifferenz regrediert (Tabelle 74, oben), ergibt sich ein positives Vorzeichen bei Lead1 der Notendifferenz. Das heißt, mit zunehmender Notendifferenz zwischen den Wahlmöglichkeiten steigt der Anteil Prüfungen in Wirtschafts- und Sozialpsychologie (gegenüber Jura und Wirtschafts- und Sozialgeschichte) im folgenden Jahr, während der Anteil Prüfungen in Jura (gegenüber Wirtschafts- und Sozialpsychologie sowie -geschichte) sinkt. Dieses Ergebnis ändert sich auch nicht, wenn die Wirtschafts-und Sozialgeschichte aus der Berechnung der Anteile wegfällt und nur das Verhältnis der Prüfungen in den beiden übrigen Wahloptionen auf deren Notendifferenz regrediert wird. Lediglich der Wert des Koeffizienten erhöht sich, was bedeutet, dass die Abhängigkeit der beiden Zeitreihen ohne die ‚Verunreinigung' des dritten Teilprüfungsgebietes noch deutlicher hervortritt.

Tabelle 73: VAR Anteil Prüfungen in Jura und Notendifferenz zwischen Jura und Wirtschafts- und Sozialpsychologie

AV: D.Anteil Prüfungen Jura	Koeffizient	Standardfehler	P>|t|
D.Anteil Prüfungen Jura (Lead1)	0.204	0.180	0.255
D.Anteil Prüfungen Jura (Lead2)	0.299	0.175	0.087
D.Notendifferenz Jura-WiSoPsy (Lead1)	**-0.085**	**0.040**	**0.035**
D.Notendifferenz Jura-WiSoPsy (Lead2)	-0.009	0.035	0.805
			n=24 (1973-1996); r^2adj=0.23
AV: D.Notendifferenz Jura-WiSoPsy			
D.Anteil Prüfungen Jura (Lead1)	1.249	0.829	0.132
D.Anteil Prüfungen Jura (Lead2)	1.058	0.806	0.190
D.Notendifferenz Jura-WiSoPsy (Lead1)	-0.291	0.187	0.119
D.Notendifferenz Jura-WiSoPsy (Lead2)	-0.181	0.163	0.265
			n=24 (1973-1996); r^2adj=0.21

Tabelle 74: VAR Anteil Prüfungen in WiSo-Psychologie und Notendifferenz zwischen Jura und WiSo-Psychologie

AV: D.Anteil Prüfungen WiSoPsy	Koeffizient	Standardfehler	P>|t|
D.Anteil Prüfungen WiSoPsy (Lead1)	0.145	0.166	0.383
D.Anteil Prüfungen WiSoPsy (Lead2)	0.382	0.155	0.014
D.Anteil Prüfungen WiSoPsy (Lead3)	0.243	0.159	0.126
D.Anteil Prüfungen WiSoPsy (Lead4)	-0.435	0.142	0.002
D.Notendifferenz Jura-WiSoPsy (Lead1)	**0.180**	**0.048**	**0.000**
D.Notendifferenz Jura-WiSoPsy (Lead2)	0.056	0.049	0.249
D.Notendifferenz Jura-WiSoPsy (Lead3)	0.013	0.045	0.774
D.Notendifferenz Jura-WiSoPsy (Lead4)	0.041	0.036	0.261
			n=22 (1975-1996); r^2adj=0.57
AV: D.Notendifferenz Jura-WiSoPsy			
D.Anteil Prüfungen WiSoPsy (Lead1)	-1.127	0.598	0.059
D.Anteil Prüfungen WiSoPsy (Lead2)	-2.244	0.558	0.000
D.Anteil Prüfungen WiSoPsy (Lead3)	0.782	0.570	0.170
D.Anteil Prüfungen WiSoPsy (Lead4)	1.929	0.512	0.000
D.Notendifferenz Jura-WiSoPsy (Lead1)	-0.418	0.171	0.014
D.Notendifferenz Jura-WiSoPsy (Lead2)	-0.141	0.175	0.422
D.Notendifferenz Jura-WiSoPsy (Lead3)	0.055	0.161	0.733
D.Notendifferenz Jura-WiSoPsy (Lead4)	0.102	0.130	0.430
			n=22 (1975-1996); r^2adj=0.62

In Bezug auf fachexterne Wahlmöglichkeiten passt jedoch nur diese eine Entwicklung zur Annahme taktisch bedingter Selbstselektionsprozesse in Teilprüfungen mit besseren Noten im Diplom. Die Höhe der Notendifferenz zwischen Wirtschafts- und Sozialgeschichte und Wirtschafts- und Sozialpsychologie zeigt keinen Einfluss auf die Prüfungsanteile Ersterer, während für die ersten Differenzen der Anteile Letzterer ein stabiles Modell mit fünf Lags einen signifikanten Einfluss des fünften Lags anzeigt. Zwar stimmt das Vorzeichen mit den Erwartungen überein, es lässt sich jedoch nicht sinnvoll begründen, warum Veränderungen in der Notendifferenz zwischen den beiden Prüfungsgebieten fünf Jahre später Einfluss auf die Wahl des Prüfungsgebietes haben sollten, in den vier näher zurückliegenden Jahren jedoch nicht. In Bezug auf fachinterne Wahlmöglichkeiten konnte nur ein einziger Fall (ebenfalls BWL an der Universität Göttingen) überprüft werden, hier zeigt sich allerdings keine passende Entwicklung der Prüfungsanteile. Eine zunehmende Selbstselektion in Teilprüfungen mit besseren Noten ist für Diplomstudiengänge demnach nicht nachzuweisen.

Abbildung 143: Durchschnittsnoten in Wahlfächern in Mathematik Diplom - TU Braunschweig (LOWESS 0.4)

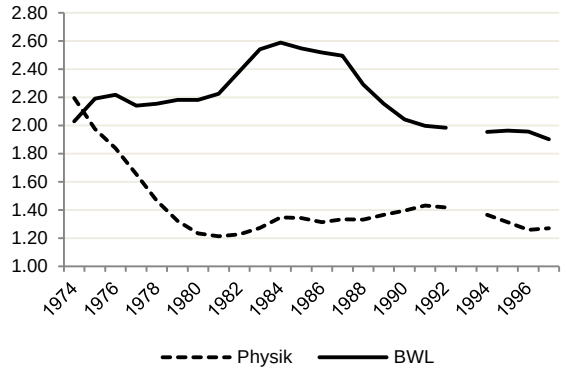


Abbildung 144: Verteilung der Prüflinge auf Wahlfächer in Mathematik Diplom - TU Braunschweig (LOWESS 0.4)

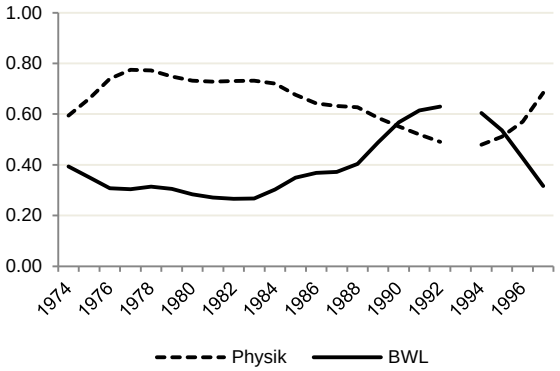


Abbildung 145: Durchschnittsnoten in Wahlfächern in Mathematik Diplom - KIT Karlsruhe (LOWESS 0.4)

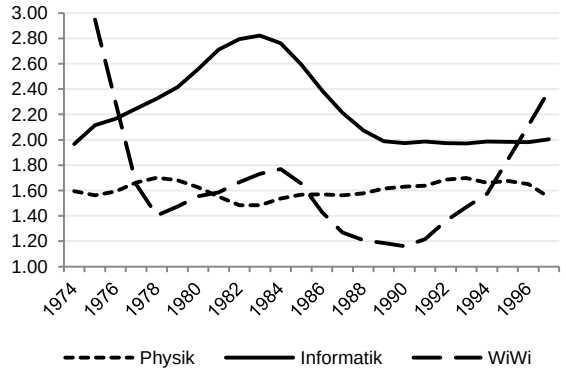


Abbildung 146: Verteilung der Prüflinge auf Wahlfächer in Mathematik Diplom - KIT Karlsruhe (LOWESS 0.4)

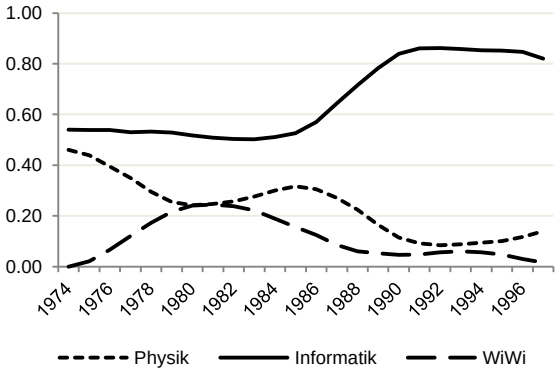


Abbildung 147: Durchschnittsnoten in Wahlfächern in Mathematik Diplom - FU Berlin (LOWESS 0.4)

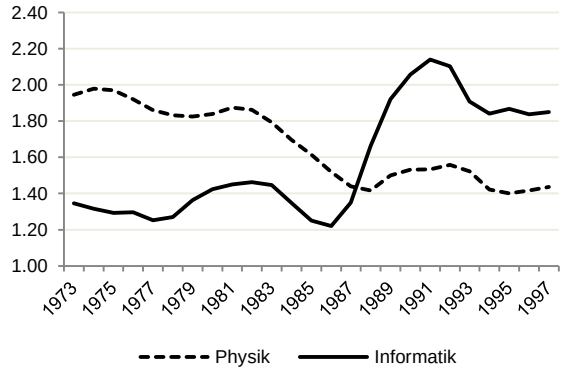


Abbildung 148: Verteilung der Prüflinge auf Wahlfächer in Mathematik Diplom - FU Berlin (LOWESS 0.4)

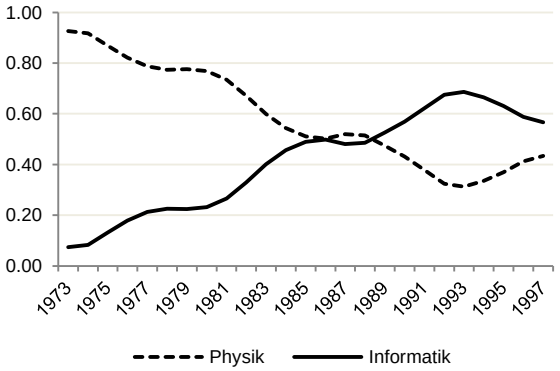


Abbildung 149: Durchschnittsnoten in Wahlfächern in Mathematik Diplom - Universität Münster 1972-1984 (LOWESS 0.4)

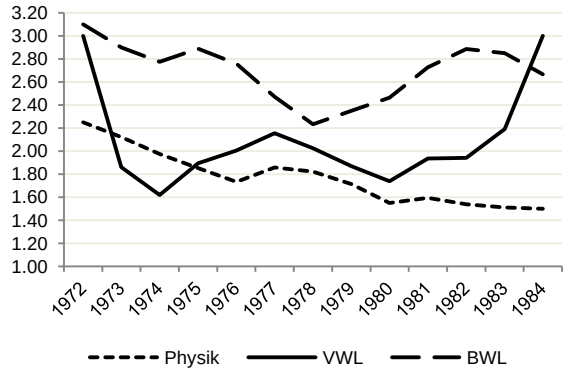


Abbildung 150: Verteilung der Prüflinge auf Wahlfächer in Mathematik Diplom - Universität Münster 1972-1984 (LOWESS 0.4)

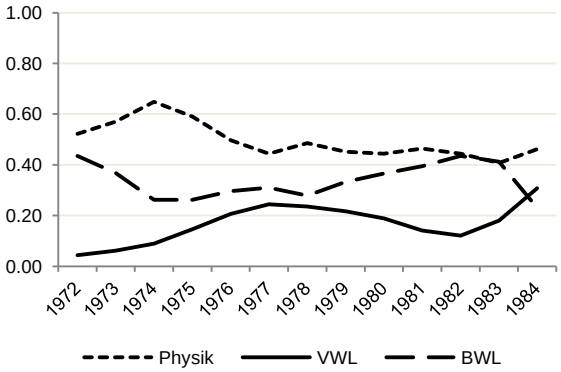


Abbildung 151: Durchschnittsnoten in Wahlfächern in Mathematik Diplom - Universität Münster 1984-1997 (LOWESS 0.4)

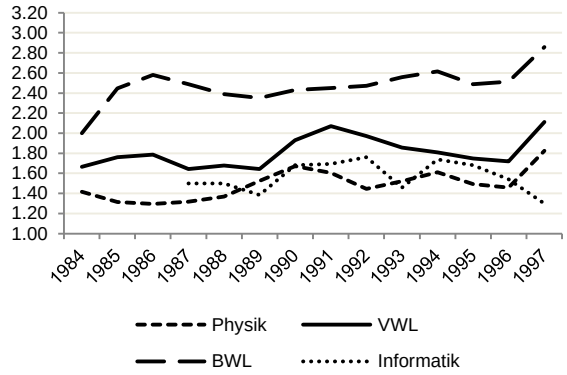


Abbildung 152: Verteilung der Prüflinge auf Wahlfächer in Mathematik Diplom - Universität Münster 1984-1997 (LOWESS 0.4)

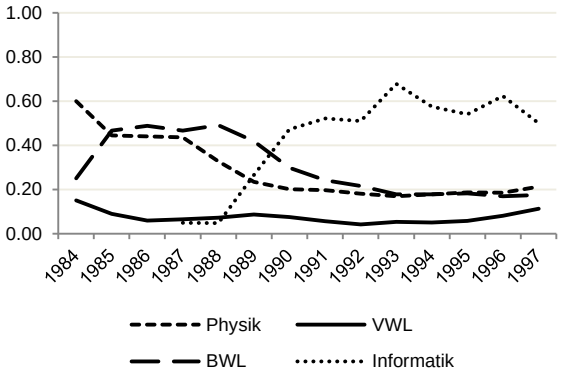


Abbildung 153: Durchschnittsnoten in Wahlfächern in Biologie Diplom - TU Braunschweig (LOWESS 0.4)

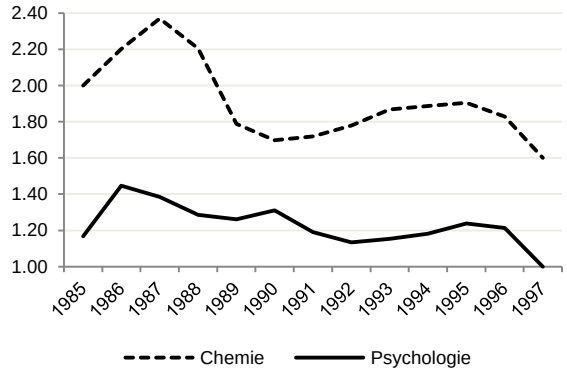


Abbildung 154: Verteilung der Prüflinge auf Wahlfächer in Biologie Diplom - TU Braunschweig (LOWESS 0.4)

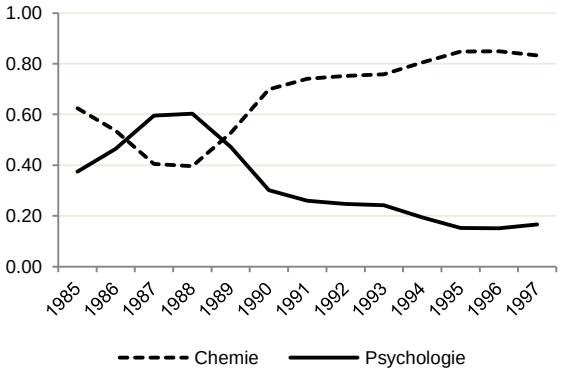


Abbildung 155: Durchschnittsnoten in Wahlfächern in VWL Diplom - FU Berlin (LOWESS 0.6)

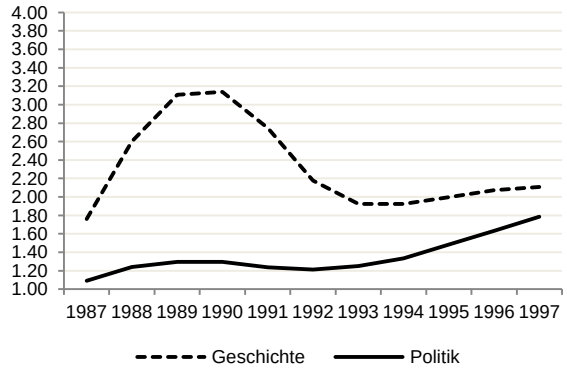


Abbildung 156: Verteilung der Prüflinge auf Wahlfächer in VWL Diplom - FU Berlin (LOWESS 0.6)

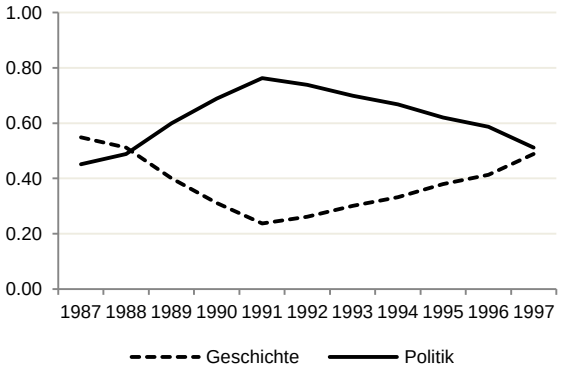


Abbildung 157: Durchschnittsnoten in Wahlfächern (intern) in BWL Diplom - Universität Göttingen (LOWESS 0.4)

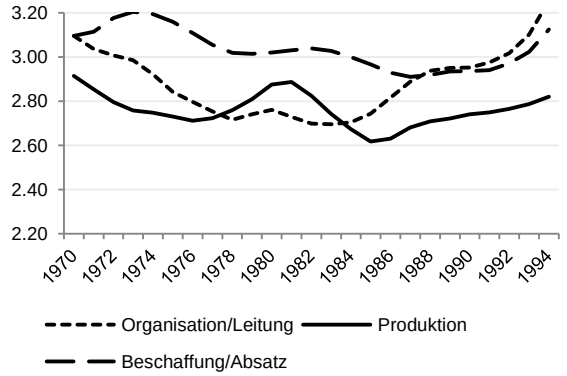


Abbildung 158: Verteilung der Prüflinge auf Wahlfächer (intern) in BWL Diplom - Universität Göttingen (LOWESS 0.4)

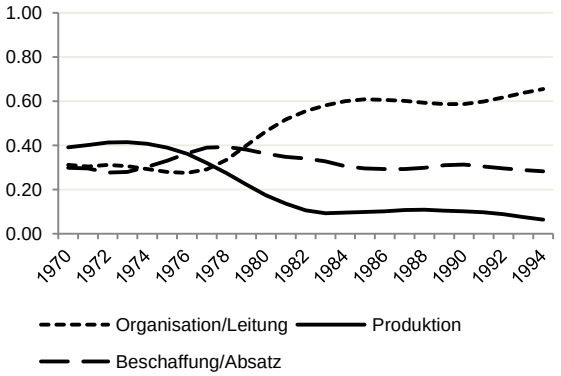


Abbildung 159: Durchschnittsnoten in Wahlfächern (extern) in BWL Diplom - Universität Göttingen (LOWESS 0.4)

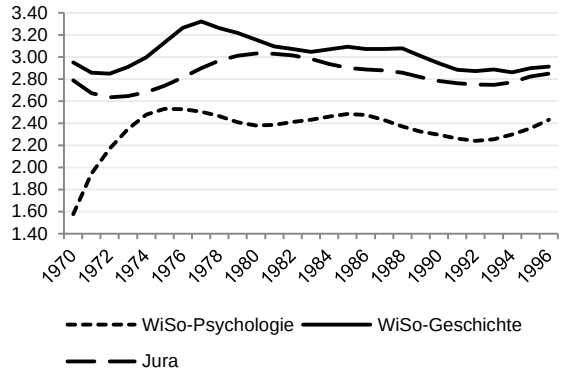


Abbildung 160: Verteilung der Prüflinge auf Wahlfächer (extern) in BWL Diplom - Universität Göttingen (LOWESS 0.4)

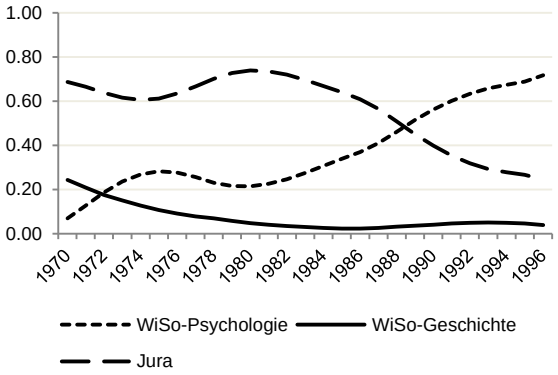


Abbildung 161: Durchschnittsnoten in Wahlfächern in Psychologie Diplom - Universität Tübingen (LOWESS 0.4)

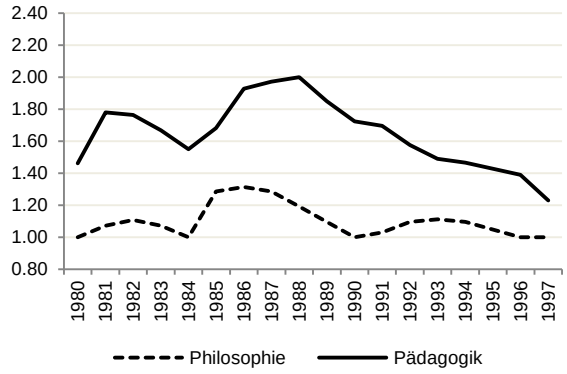
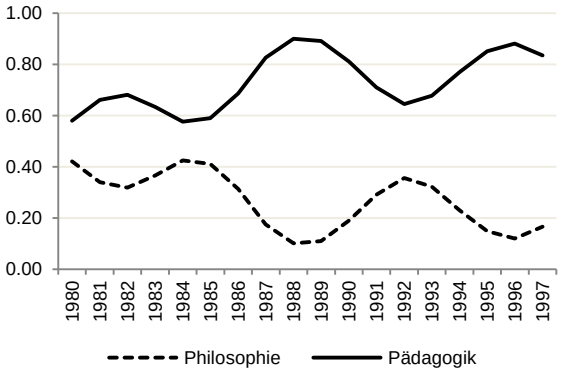


Abbildung 162: Verteilung der Prüflinge auf Wahlfächer in Psychologie Diplom - Universität Tübingen (LOWESS 0.4)



Zusammenfassung

Das vorliegende Datenmaterial ist je nach potentiellem Einflussfaktor in unterschiedlichem Maße geeignet, leistungskonforme Effekte auf die Notengebung zu überprüfen. Für die kompositionellen Studierendenmerkmale *Anteil Teilzeitstudierende*, *Stipendiat*innenanteil*, und *soziale Herkunft* können nur Modellrechnungen aufgestellt werden, die den maximalen Einfluss unter angenommenen bzw. aus den dem Konstanzer Studierendensurvey entnommenen Zwischenprüfungsergebnissen ermittelten Notenunterschieden darstellen. Diese zudem meist auf die Studiengangebene beschränkten Rechnungen genügen jedoch, um zu zeigen, dass lediglich die soziale Herkunft als relevante Einflussgröße zur Erklärung von Notenniveauunterschieden zwischen Studiengängen geeignet ist. Hier ist allerdings zu beachten, dass dies nur für bestimmte Paarvergleiche gilt.

Auch Im Zeitverlauf hat von diesen vier Variablen nur die soziale Herkunft überhaupt einen Wandel in ihrer Verteilung über die Studierenden genommen, der mit einer Verbesserung der Noten, wie sie für die meisten Studiengänge nachgewiesen ist, in (sehr!) begrenztem Maß in Einklang steht. Für einen generellen Alterseffekt findet sich ebenfalls kein Hinweis in den Daten des Konstanzer Studierendensurveys, allerdings ist studiengangspezifisch nicht auszuschließen, dass es einen notenverschlechternden Einfluss der Studiendauer gibt, der das Notenniveau in einzelnen Studiengängen tendenziell erhöht. Auch hier wäre eine Erklärung von Notendifferenzen allerdings auf einzelne Paarver-

gleiche beschränkt und es kann nicht von großer Erklärungskraft ausgegangen werden. Ein Einfluss auf die zeitliche Entwicklung ist hier ebenfalls auszuschließen.

Der Geschlechtskomposition kommt ebenso wie der Zusammensetzung nach sozialer Herkunft eine begrenzte paarvergleichsspezifische Erklärungskraft für Unterschiede im Notenniveau zu - sowohl auf Studiengang- wie auch auf Hochschulebene. Im Längsschnitt zeigen sich auf Studiengangebene vereinzelt kurze Phasen, in denen die Erhöhung des Frauenanteils das Notenniveau senkt. Diese vereinzelten Episoden reichen aber nicht aus, um die festgestellten langfristigen und in Zyklen kontinuierlich verlaufenden Verbesserungsprozesse erklären zu können. Und auch auf Hochschulebene lässt sich nicht von einem systematischen Einfluss der Entwicklung der Geschlechtskomposition auf die Notenentwicklung sprechen. Dass die Geschlechterkomposition nur unsystematischen Einfluss besitzt, zeigt sich auch bei Grözinger (2017).

Für die Selbstselektionsthese lassen sich keinerlei überzeugende Hinweise finden - weder auf Studiengangs- oder Hochschul- noch auf Teilprüfungsebene. Gegen die Erklärungskraft der These lässt sich zudem anführen, dass eine Abwanderung in besser benotete Alternativen, egal auf welcher Ebene, vor allem leistungsschwächere Studierende umfassen dürfte - woraufhin sich das Notenniveau in diesen Studiengängen, Hochschulen oder Teilprüfungsmodulen wiederum verschlechtern, in den von Abwanderung betroffenen hingegen verbessern müsste. Die Folge wäre also eher eine Angleichung der Noten, keine allgemeine Verbesserung.

Die Eingangseignung scheint eine relevante Größe bei der Notenvergabe zu sein: Sowohl im Quer- als auch im Längsschnitt weisen die Stichprobendaten bzw. die des Konstanzer Studierendensurveys auf Studiengangebene auf gerechtfertigte Notenverbesserungen bei verbesserten Eingangseignungen hin[100]. Auf Hochschulebene finden sich im Querschnitt allerdings keine entsprechenden Hinweise, im Längsschnitt zeigen sich gemischte Ergebnisse. FH2b (Zeitlich stabile Differenzen im Notenniveau zwischen fachlich abgegrenzten Studiengängen lassen sich zum Teil auf fachspezifische Prüfungsbedingungen zurückführen) und FH3b (Fach-/studiengangspezifische Notenverläufe lassen sich zum Teil auf fachspezifische Entwicklungen der Prüfungsbedingungen zurückführen) lassen sich damit bestätigen.

8.2.2 Leistungsunabhängige Prüfungsbedingungen

Forschungsintensität

Angenommen, Prüfer*innen würden Prüflingen bei einer Präferenz für die Forschung bessere Noten erteilen, um weniger Zeit für die Prüfungsvorbereitung und die Verhandlung von Beschwerden nut-

[100] Dass die Eingangseignung in Relation zu den anderen bis hier behandelten Prüfungsbedingungen eine gewichtigere Rolle spielt, zeigt auch eine multivariate Regression der im Konstanzer Studierendensurvey enthaltenen Zwischenprüfungsnote auf die dort enthaltenen leistungskonformen Einflussfaktoren. Die Abiturnote weist dort den mit Abstand größten beta-Wert auf (siehe Anhang: Tab.A21).

zen zu müssen, sollten Unterschiede in der Forschungsintensität mit Unterschieden im Notenniveau einhergehen (vgl. Franz 2010). Die Forschungsintensität zwischen Fächern bzw. Studiengängen anhand der klassischen Indikatoren (Ausgaben für Forschung und Entwicklung, Anzahl Promotionen, Anzahl Publikationen) zu vergleichen scheint wenig sinnvoll, da zum Teil völlig unterschiedliche epistemologische Voraussetzungen für Forschung existieren.

Auch ein Versuch Grözingers (2017), die Forschungsintensität anhand der Relation drittmittelbeschäftigter Mitarbeiter*innen zur Anzahl Professor*innen zu erfassen, da der Anteil der Drittmittelangestellten hoch mit der Forschungsintensität korrelieren sollte, bringt keine eindeutigen Ergebnisse hervor. Möglicherweise könnte dieses Vorgehen auf den Vergleich zwischen Hochschulen innerhalb eines Studiengangs angewendet werden. Wissenschaftliche Mitarbeiter*innen können jedoch sowohl aus Haushaltsmitteln oder Drittmitteln als auch aus beiden Quellen finanziert sein und die absolute Anzahl der Stellen gibt keinen Aufschluss darüber, ob es sich um ganze, halbe oder anderweitig anteilige Stellen handelt.

Eine Berechnung dieses Indikators auf Hochschulebene würde daher zu ungenau ausfallen, um aus einem rein deskriptiven Vergleich zwischen den wenigen Hochschulen in dieser Stichprobe einen Rückschluss auf die dort herrschenden Differenzen im Notenniveau ziehen zu können. Da auf Hochschulebene ebenfalls keine Informationen zu den oben genannten klassischen Indikatoren vorliegen, bleibt die Frage nach einem möglichen Einfluss der Forschungsintensität auf die Noten auf Hochschulebene an dieser Stelle offen. Auf Fachebene lässt sich an dieser Stelle auf die konsistenten Befunde der Fachkulturforschung (siehe Kapitel 3.2) verweisen, nach denen der Forschungsaufwand sowie die Präferenz für Forschung gegenüber Lehre in den Natur- und Ingenieurswissenschaften am höchsten, in den Geistes- und Sozialwissenschaften am niedrigsten ist. In Kombination mit dem Ergebnis, dass die naturwissenschaftlichen Studiengänge im sample die besten Noten vergeben, steht dies zumindest in Einklang mit der Annahme eines Zusammenhangs zwischen guten Noten und hoher Forschungsintensität.

Zusammensetzung der Lehrenden

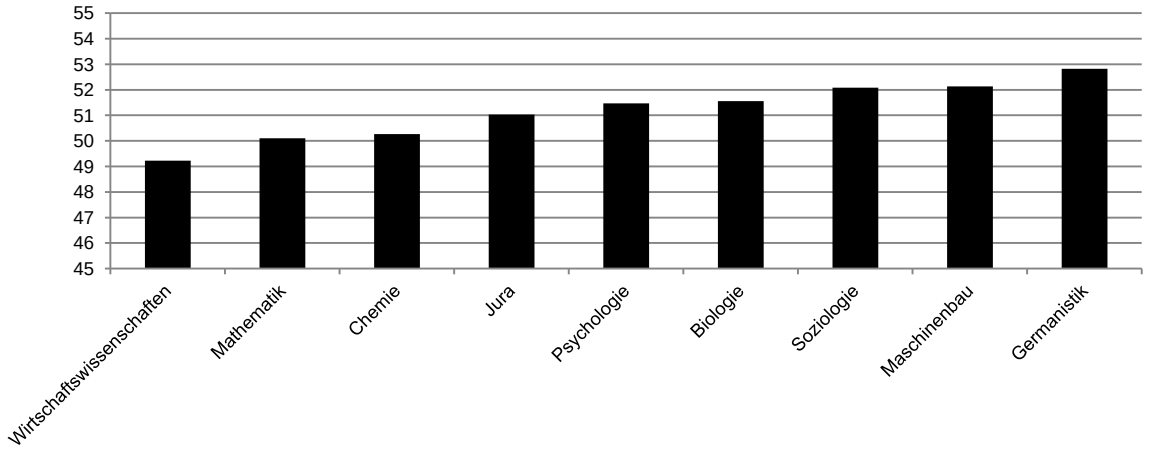
Ebenso wie die Zusammensetzung der Studierenden könnte die Zusammensetzung der Lehrenden einen Einfluss auf das Notenniveau besitzen, sollten sich die Lehrkörper zwischen oder innerhalb von Studiengängen systematisch hinsichtlich bestimmter Merkmale unterscheiden, die mit der Notenvergabe zusammenhängen. Im Gegensatz zur Studierendenkomposition wäre dies jedoch eine leistungsunabhängige Verzerrung der Notenvergabe. Während zur sozialen Herkunft, die möglicherweise zu diesen Merkmalen zählt ebenso wie zu den politischen Einstellungen der Lehrenden keine ausreichenden Daten zur Verfügung stehen, um Studiengang- geschweige denn Hochschulvergleiche

durchzuführen , die dem sample entsprechen[101], bietet die Hochschulpersonalstatistik zumindest auf Fachebene (STB=Studienbereich, ohne Lehramt) aussagekräftige Daten zum Alter der Professor*innen und zum Professorinnenanteil.

Abbildung 163 zeigt das Durchschnittsalter der Professor*innen nach Fachbereich für den Aggregationszeitraum von 1995-2013. Die Spannweite zwischen dem jüngsten Altersdurchschnitt in den Wirtschaftswissenschaften und dem ältesten in Germanistik beträgt R=3.6 Jahre - diese minimale Differenz wird kaum zu Unterschieden in der Notenvergabe führen, die grundsätzlich auf das Alter (oder die Prüfungserfahrung) reagiert (bessere Noten bei steigendem Alter zeigen sich bei Grözinger (2017) in allen Fachbereichen die den sample Studiengängen entsprechen). Dass sich die Altersstruktur der Professor*innen zwischen Fächergruppen nicht unterscheidet, können Lundgreen et al. (2009) und der Wissenschaftsrat (1998b) zeigen.

Da im Querschnitt keine nennenswerten Unterschiede existieren, genügen die vorhandenen fach(gruppen)übergreifende Daten, um den möglichen Einfluss der Altersstruktur im Längsschnitt zu behandeln: Hier geben die Daten mehrerer Erhebungszeitpunkte seit 1960 keine Hinweise auf einen langfristigen Trend zu kleiner oder größer werdenden Durchschnitten (Wissenschaftsrat 1988; 1995; 1998b). Die Daten auf Hochschulebene zeigen für die sample Universitäten ebenfalls nur geringfügige Unterschiede im Altersdurchschnitt von 3-4 Jahren[102]. Aufgrund der (mindestens) facheinheitlichen Ausbildungswege in den letzten Jahrzehnten ist dies folgerichtig. Dass bestimmte Hochschulen im Durchschnitt systematisch jüngere oder ältere Professor*innen beheimaten war bisher nicht wahrscheinlich. Dies könnte sich in Zukunft durch unterschiedliche Umsetzungen von neuen, alternativen Karrierewegen wie der Juniorprofessur oder der Tenure-Track-Professur allerdings ändern.

Abbildung 163: Durchschnittsalter der Professor*innen nach Fachbereich (STB) 1995-2013



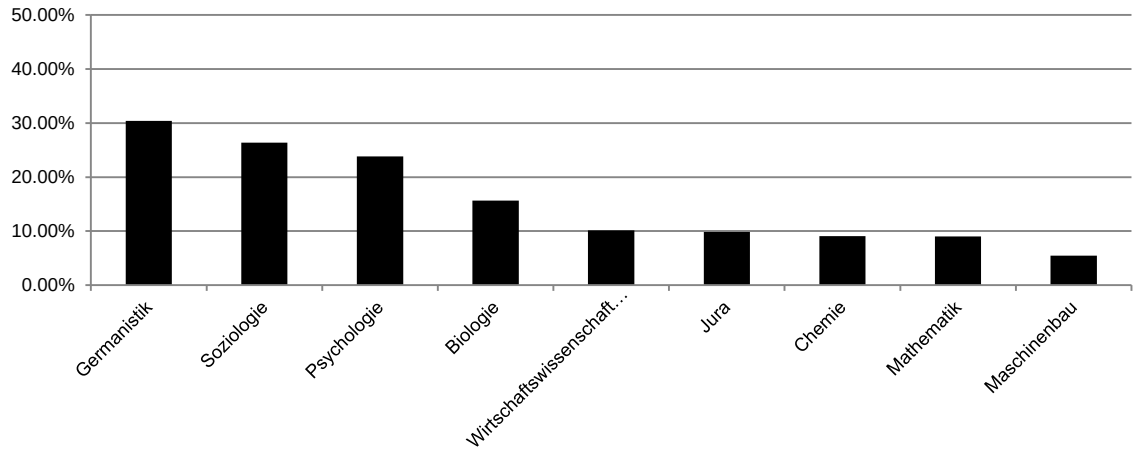
Quelle: FDZ des Statistischen Landesamts Schleswig-Holstein, Berechnung durch Prof. Dr. Gerd Grözinger – keine Fallzahlen verfügbar, da die Hochschulpersonalstatistik im verwendeten Datensatz den Fällen der Hochschulprüfungsstatistik zugespielt wurde.

[101] Möller (2015:229) präsentiert eine Abstufung nordrhein-westfälischer Professor*innen nach sozialer Herkunft, der zufolge Jurist*innen, Mediziner*innen und Sportwissenschaftler*innen am häufigsten aus gehobenen Verhältnissen stammen, Psycholog*innen/Pädagog*innen und Forstwissenschaftler*innen am seltensten.

[102] Aus Platzgründen nicht dargestellt. Die Daten können bei Interesse gerne beim Autor angefragt werden.

Hinsichtlich des Professorinnenanteils zeigen sich fachspezifisch deutliche Unterschiede, die ähnlich bereits bekannt sind (etwa Möller 2015): Den größten Anteil Professorinnen unter den im sample vertretenen Fachbereichen weist die Germanistik mit 30.4% auf, den geringsten der Maschinenbau mit 5.4% im selben Aggregationszeitraum - ein Unterschied von 25 Prozentpunkten (Abb.164). Da Grözinger (2017) zeigen kann, dass mit steigendem Professorinnenanteil die Noten sinken (Ausnahme Wirtschaftswissenschaften), ist dies ein bedeutsamer Unterschied. So fällt dann auch der vergleichsweise hohe Anteil in Psychologie ins Auge - hier würde ein geringerer Abstand zu den nachfolgenden Fachbereichen die Notendifferenz ein wenig verringern. Wie auch für die Kompositionsmerkmale der Studierenden fallen die Auswirkungen auf die Unterschiede im Notenniveau jedoch paarvergleichsspezifisch aus, da die Abstufung der Anteile nicht der Rangfolge der Notenhierarchie entspricht.

Abbildung 164: Anteil der Professorinnen nach Fachbereich (STB) 1995-2013



Quelle: FDZ des Statistischen Landesamts Schleswig-Holstein, Berechnung durch Prof. Dr. Gerd Grözinger – keine Fallzahlen verfügbar, da die Hochschulpersonalstatistik im verwendeten Datensatz den Fällen der Hochschulprüfungsstatistik zugespielt wurde.

Da Frauen in der Wissenschaft noch nicht lange über ähnliche Karrierechancen wie Männer verfügen (falls sie das inzwischen überhaupt tun), ist davon auszugehen, dass auch die hier dargestellten Querschnittsunterschiede in diesem Maße noch nicht lange existieren, einfach, weil erst in jüngerer Vergangenheit überhaupt genug Frauen Professorinnen werden konnten, um einen relevanten Anteil im Fachbereich stellen zu können. Entsprechend zeigen die Daten von Lundgreen et al. (2009:67), dass sich auf Fachgruppenebene erst ab Anfang der 1990er Jahre Unterschiede in der Geschlechterkomposition entwickeln. Die langfristige Notenverbesserung kann durch einen steigenden Anteil an Professorinnen also nicht erklärt werden - zumal dort, wo die Anteile am stärksten gestiegen sind (ebd.: hier auch fachspezifisch angegeben), nämlich in Germanistik und Psychologie, keine langfristige Verbesserung auf Studienganglevel festzustellen ist (Germanistik) bzw. die Verbesserung Anfang der 1990er schon lange abgeschlossen war (Psychologie).

Auch Hochschulebene sind leider zu häufig zu wenige Professorinnen an den Hochschulen des samples angestellt (gewesen), als dass ein Vergleich sinnvoll wäre.

Unterschiedliche Notenniveaus

Sollten Verbesserungsprozesse im Notenniveau nicht unabhängig von Unterschieden in der Höhe des Notenniveaus zwischen Fächern und Hochschulen ablaufen, etwa weil die im Querschnitt nachgewiesenen Unterschiede im Notenniveau Prüfende mit vergleichsweise schlechten Notenniveaus dazu veranlassen, bessere Noten zu vergeben, um ihre Prüflinge im Vergleich mit besser bewerteten Studierenden nicht zu benachteiligen, müsste sich dies in Zusammenhängen der Notenreihen zeigen. Die Zeitreihe der schlechteren Noten müsste sich prinzipiell der besseren Zeitreihe angleichen.

Da jedoch im Zweifelsfall auch die bessere Reihe einer langfristigen Notenverbesserung unterliegt, würde eine Spiralentwicklung in Gang gesetzt werden, die sich in einer kurzfristigen zeitverzögerten Reaktion der schlechteren Noten auf die Entwicklungen der besseren Noten zeigen müsste. Das Argument der Nicht-Benachteiligung auf dem Arbeitsmarkt ist jedoch nur für die Entwicklung hochschulspezifischer Noten innerhalb von Studiengängen plausibel (solange nicht für einen gemeinsamen Arbeitsmarkt produziert wird). Prüfer*innen in VWL werden kaum ihre Noten an die besseren Noten in Chemie anpassen. Zudem müssen, damit der beschriebene Mechanismus greifen kann, über einen längeren Zeitraum merkbare Unterschiede im Notenniveau zwischen einzelnen Hochschulen herrschen. Ist in einem solchen Zeitraum eine Verbesserung der Hochschule mit den schlechteren Noten gegeben, steht diese Entwicklung in Einklang mit der These. Dies ist im sample in 36 Paarvergleichen der Fall. VAR Modelle zeigen allerdings, dass nur in zwei (Originaldaten) bzw. vier (geglättete Daten) Paarvergleichen ein positiver Zusammenhang zwischen den Notenreihen besteht, der ein Jahr Reaktionszeit aufweist. Die reine Existenz von Niveauunterschieden beeinflusst die Entwicklung der Noten also nicht. Dies ist nachvollziehbar, müssten die Prüfenden doch langfristig einen Überblick über die Rangfolge der in ihrem Studiengang vergebenen Noten an den einzelnen Hochschulen im Blick haben, um ihre eigenen Noten einzuordnen und auf die der anderen Standorte reagieren zu können. Naheliegender ist die Vermutung, dass Prüfende ihre Noten hochschulintern an die gegebenen Niveaus anpassen, die sie unmittelbar mitbekommen.

Formale Prüfungsbedingungen und eingesetzte Prüfungsverfahren

Als gewichtige Einflüsse auf das Prüfungsergebnis sollten die per Prüfungsordnung festgelegten formalen Prüfungsbedingungen wirken. Neben Möglichkeiten zur Vermeidung von schlechten Noten etwa durch *Freiversuche* (Birnbaum 1977) fallen in diesen Bereich mehrere mögliche Faktoren:

Eine unterschiedliche *Anzahl der Teilprüfungsleistungen* dürfte mit einer unterschiedlich hohen Gesamtbelastung des Prüflings einhergehen, weshalb auch mit unterschiedlichen Leistungen zu rechnen ist: Je mehr Einzelleistungen, umso schlechter dürfte die Note sein. Ein hoher *Anteil an Nebenfachprüfungsleistungen* dürfte ebenfalls notenverschlechternd wirken, da der Fokus des Studierenden in

der Regel eher auf seinem Hauptfach als auf dem Nebenfach liegen wird. Die *Dauer*, die *für das Erbringen einzelner Prüfungsleistungen* festgelegt ist, sollte ebenfalls einen Einfluss aufweisen. Hier sind prinzipiell zwei Wirkungsrichtungen denkbar: Entweder geht mit einer längeren Bearbeitungsdauer eine bessere Note einher, weil der Prüfling mehr Zeit zur Bearbeitung der Aufgabenstellung hat oder eine schlechtere Note, weil er entsprechend der längeren Bearbeitungsdauer auch mehr oder umfangreichere Aufgaben bewältigen muss. Auch, ob die Kandidat*innen das *Thema der wissenschaftlichen Abschlussarbeit selbst wählen* bzw. mitbestimmen dürfen oder ob eine vorgegebene Fragestellung zu bearbeiten ist, könnte in zwei gegensätzliche Richtungen wirken: Einerseits könnte die höhere intrinsische Motivation bei eigener Auswahl bessere Leistungen fördern, andererseits haben Studierende häufig Probleme eine geeignete Fragestellung für eine Abschlussarbeit zu wählen, die dem geforderten Umfang gerecht wird. Dies könnte v.a. bei Überfrachtung der Arbeit die Leistung und damit die Note negativ beeinflussen.

Schließlich bestehen bei der Bewertung der Leistungen Unterschiede im *verwendeten Notensystem* (ganze Noten, halbe Noten, Viertelnoten, Differenzierung durch 0.3/0.7) die im Zeitverlauf größer sind als etwa im Querschnitt zwischen Studiengängen (dort aber immer noch größer als zwischen Hochschulen). Hier ist es fraglich, ob eher grobe oder eher kleine Unterteilungen im Notensystem die Gesamtnote und damit den Notendurchschnitt zum Besseren oder Schlechteren hin beeinflussen: Es ließe sich argumentieren, dass bei ganzen Teilprüfungsnoten unter bestimmten Konstellationen der einzelnen Teilnoten eine bessere Gesamtnote erzielt wird, wenn die Leistungen nach einer differenzierteren Benotung zum Großteil eher im Grenzbereich zu einer schlechteren (ganzen) Note liegen. Umgekehrt ist aber auch eine bessere Note bei differenzierter benoteten Leistungen möglich, wenn diese zum Großteil im Grenzbereich zur jeweils besseren (ganzen) Note liegen. Unter der Annahme, dass sich diese beiden rechnerischen Einflüsse in einer großen Masse von Prüflingen ausmitteln, bleibt allerdings noch ein weiteres Argument, dass einen Effekt des Notensystems an die Selektionsneigung der Prüfenden koppelt: Je gröber die Unterteilung und damit der Abstand zur nächstschlechteren Note, umso überzeugter sollten die Prüfenden sein, mit ihrer Bewertung richtig zu liegen. Daraus folgt, dass gröbere Unterteilungen bei Prüfenden mit einer starken Selektionsneigung eher zu schlechteren Noten führen sollten, da sie in der Regel vermutlich das Ziel verfolgen, die besten Prüflinge von den übrigen zu trennen, während bei geringer Selektionsneigung eher noch ein Auge zugedrückt und die Leistung von Wackelkandidat*innen von einem „gut“ noch zu einem „sehr gut“ werden könnte. Zudem wird beim Übergang von ganzen zu differenzierten Noten die Möglichkeit eines Substitutionseffekts in Erwägung gezogen, in dessen Folge die Anzahl der bisher üblichen Ausprägungen einfach übernommen wird, eine ‚1.5‘ also beispielsweise die neue ‚2‘ darstellt (Suslow 1976).

Auch die eingesetzten Prüfungsverfahren dürften einen Einfluss auf die Notenhöhe besitzen. Wie in Tabelle 75 abzulesen, gilt für alle Studiengänge (mit Ausnahme von Diplom Soziologie an der FU Ber-

lin), dass das durchschnittliche Notenniveau in der schriftlichen Abschlussarbeit unter dem der mündlichen Prüfungen liegt. In Klausuren werden die schlechtesten Ergebnisse erzielt[103] (in Biologie und Chemie wurden an den im sample enthaltenen Hochschulen im betrachteten Zeitraum keine Klausuren in der Diplom-Abschlussprüfung geschrieben, in Mathematik nur in wenigen Jahren und meist als optionale Prüfungsform. In VWL und BWL sowie in den beiden Magisterstudiengängen konnten die Einzelresultate der mündlichen Prüfungen und der Klausuren leider nur vereinzelt voneinander getrennt erhoben werden). Daher geht auch das Verhältnis der Hausarbeit zur Anzahl mündlicher Prüfungen sowie das Verhältnis der Klausur(en) zur Anzahl mündlicher Prüfungen in die Analysen ein. Zudem wird die Gewichtung der Hausarbeit (die in der Regel höher liegt als die der übrigen Einzelteilleistungen) als Prüfungsform mit dem besten Notenniveau berücksichtigt.

Tabelle 75: Notenniveaus (aggregiert 1960-1997) in unterschiedlichen Prüfungsverfahren nach Studiengang

Studiengang	Abschlussarbeit (n)	Mündliche Prüfungen (n)	Teilprüfungen mündlich + Klausur (n)	Klausur/en (n)
Mathematik Diplom	x̄=1.59, s=0.78 (6 846)	x̄=1.76, s=0.71 (6 833)	--	--
Mathematik Lehramt	x̄=1.69, s=0.81 (4 995)	--	x̄=2.28, s=0.81 (6 699)	--
Chemie Diplom	x̄=1.31, s=0.53 (12 395)	x̄=1.92, s=0.73 (12 395)	--	--
Biologie Diplom	x̄=1.38, s=0.62 (11 635)	x̄=1.50, s=0.52 (11 621)	--	--
Psychologie Diplom	x̄=1.51, s=0.72 (11 518)	x̄=1.63, s=0.51 (11 850)	--	x̄=1.80, s=0.79 (11 101)
VWL Diplom	x̄=2.35, s=0.93 (11 822)	--	x̄=2.72, s=0.68 (11 801)	--
BWL Diplom	x̄=2.52, s=0.90 (26 648)	--	x̄=2.85, s=0.63 (26 616)	--
Maschinenbau Diplom[a]	x̄=1.28, s=0.52 (12 691)	--	--	--
Soziologie Diplom[b]	x̄=1.65, s=0.67 (3 094)	x̄=1.53, s=0.48 (3 092)	--	--
Soziologie Magister	x̄=1.97, s=0.83 (1 429)	x̄=2.02, s=0.67 (163)	x̄=2.09, s=0.70 (1 258)	x̄=2.09, s=0.80 (163)
Germanistik Magister	x̄=1.93, s=0.85 (6 292)	x̄=2.07, s=0.73 (1 021)	x̄=1.98, s=0.70 (5 014)	x̄=2.34, s=1.76 (1 021)
Deutsch Lehramt	x̄=1.96, s=0.86 (13 015)	--	x̄=2.41, s=0.75 (16 684)	--

[a]für Maschinenbau Diplom lässt sich aufgrund der flexiblen Gestaltungsmöglichkeiten in der modularen Teilprüfungsbelegung nicht nachvollziehen, welche Prüfungen nur mündlich, nur schriftlich oder kombiniert durchgeführt wurden, [b]nur FU Berlin

Den einzelnen Prüflingen konnten bei der Erhebung zum Großteil (n=86 135) die gültigen Prüfungsordnungen zugeordnet werden. Prinzipiell sind damit zu allen genannten potentiellen Einflussfaktoren Informationen bis einschließlich des Prüfungsjahrgangs 1997 vorhanden. Die für Klausuren vorgesehene Bearbeitungsdauer unterscheidet sich zwischen den Diplomstudiengängen, in denen Klausuren gestellt wurden (VWL, BWL, Psychologie) allerdings nicht, weshalb diese Variable als Einflussgröße bereits ausgeschlossen werden kann. Die in den Prüfungsordnungen für mündliche Prüfungen angegebene Bearbeitungsdauer umfasst zum einen häufig Zeitspannen statt konkreten Werten. Und auch eine Mittelwertbildung der Zeitspannen o.ä. stellt keine sinnvolle Operationalisierung dar, da in einigen Prüfungsordnungen getrennte, in anderen gemeinsame Vorgaben für Haupt- und Nebenfächer vorliegen und ein Einfluss auf die Gesamtnote damit nicht vergleichbar erfasst werden kann. Eine Aufnahme in ein multiples Regressionsmodell, das für die übrigen Variablen berechnet werden kann, scheidet damit für diese beiden Variablen ebenso aus, wie für die Größe ‚Freiversuch', da Frei-

[103] Welche Prüfungsform die erbrachte Leistung am genauesten abbildet, kann nur anhand der Notenergebnisse an dieser Stelle genauso wenig überprüft werden, wie die eingangs geäußerte Vermutung, dass sich Prüfungsleistungen im Zeitverlauf aufgrund verbesserter Testinstrumente genauer erfassen lassen. Dass die Resultate mündlicher Prüfungen stärker streuen als die Ergebnisse von Klausuren trifft nicht zu.

versuche an den meisten deutschen Hochschulen erst im Laufe der 1990er Jahre eingeführt wurden, was zu einer extrem niedrigen Zahl an Prüflingen führt, für die dieses Merkmal eine positive Ausprägung aufweist.

Sinnvoll ist ein Vergleich der formalen Prüfungsbedingungen nur innerhalb einer Abschlussart. Aus diesem Grund werden die Lehramts- und Magisterstudiengänge, ebenso wie das Staatsexamen Jura aus den Analysen ausgeschlossen. Auch Maschinenbau wird nicht berücksichtigt, da aufgrund des modularen Prüfungssystems zu viel Gestaltungsspielraum in den Prüfungsvorgaben herrscht.

Mathematik Diplom

Für alle Studiengänge gilt, dass Informationen über die anhand der jeweiligen Prüfungsordnung zugewiesenen formalen Prüfungsbedingungen und eingesetzten Prüfungsverfahren in geringerem Umfang zur Verfügung stehen als es bei den Variablen ‚Anzahl der Teilprüfungen', ‚Verhältnis Nebenfach zu Hauptfach' und ‚Verhältnis Hausarbeit zu mündlichen Prüfungen' der Fall ist, welche direkt aus den erhobenen Daten berechnet werden können, womit Werte für (fast) alle Fälle vorliegen. Aus diesem Grund werden im Folgenden jeweils zwei Regressionsmodelle berechnet, die einmal alle Fälle, aber nur die vollständig erfassten Variablen, einmal alle (relevanten) Variablen, aber nur eine geringere Fallzahl enthalten.

In Mathematik reduziert sich die Anzahl brauchbarer Variablen durch zwei Besonderheiten. Zum einen sind die zwei Variablen ‚Verhältnis Klausur(en) zu mündlichen Prüfungen' und ‚Dauer der Hausarbeit' Konstanten und somit von der Varianzaufklärung der Noten ausgeschlossen. Zum anderen entsprechen sich die Variablen ‚Anzahl der Teilprüfungen', ‚Verhältnis Hausarbeit zu mündlichen Prüfungen', ‚Thema der Hausarbeit selbst wählbar', die Dummys zur Notendifferenzierung und ‚Verhältnis Nebenfach zu Hauptfach' aufgrund paralleler Änderungen in der Prüfungsordnung entweder in positiver oder negativer Form vollständig[104]. Es verbleibt damit für beide Modelle jeweils nur eine Einflussvariable zusätzlich zur Zeitvariable und den Dummys, die die Hochschulzugehörigkeit kontrollieren.

Es zeigt sich, dass mit einer höheren Nebenfachratio eine Verschlechterung der Note einhergeht (Modell A). Das Ergebnis ist allerdings nur sehr knapp signifikant und neben der allgemein geringen Varianzaufklärung (r^2=0.02) trägt der Einschluss der zusätzlichen unabhängigen Variable gerade einmal zu einer Verbesserung des r^2 von 0.001 bei. Die gleiche minimale Verbesserung bewirkt die Er-

[104] Bis auf den Studiengang Psychologie weisen alle Studiengänge eine perfekte Multikollinearität der Variablen „Anzahl der Teilprüfungen" und „Verhältnis Hausarbeit zu mündlichen Prüfungen" auf, da immer wenn sich die Anzahl der Teilprüfungen erhöht, sich auch das Verhältnis zwischen den beiden Prüfungsverfahren parallel dazu verändert, wenn die Teilprüfungen eine mündliche Prüfung darstellen oder beinhalten. Im Folgenden wird die Anzahl der Teilprüfungen deshalb nur für den Studiengang Psychologie gesondert berücksichtigt und ansonsten davon ausgegangen, dass es das Verhältnis der Hausarbeit zu den mündlichen Prüfungen ist, das Einfluss auf die Notenhöhe ausübt, da, wie gezeigt wurde, in den schriftlichen Hausarbeiten in allen Studiengängen bessere Noten erzielt werden als in den mündlichen Prüfungen.

weiterung des Modells um die Variable ‚Gewicht der Hausarbeit' (Modell B - ohne Karlsruhe und Berlin). Der Effekt ist zudem knapp nicht signifikant, wenn das Vorzeichen auch in die erwartete Richtung zeigt. Der erwartet positive Effekt der höheren Nebenfachratio ist nun hochsignifikant, durch die perfekte Multikollinearität mit den oben genannten ausgeschlossenen Variablen ist jedoch unklar, ob hier der höhere Nebenfachanteil an sich wirksam ist oder einer der anderen Einflüsse.

Chemie Diplom

Das Verhältnis Nebenfach- zu Hauptfachprüfungen weist in Chemie unter Kontrolle der Zeit und Hochschulzugehörigkeit wider Erwarten ein negatives Vorzeichen auf, ist allerdings auch nicht signifikant. Drei Mündliche Prüfungen als Gegengewicht zur schriftlichen Abschlussarbeit wirken hingegen wie erwartet notenverbessernd im Vergleich zu vier mündlichen Prüfungen. Die Güte des Modells ist mit einem r^2 von 0.09 zwar höher als in Mathematik, liegt aber immer noch niedrig.

Gegenüber den Kontrollvariablen bringen die beiden unabhängigen Variablen einen Zugewinn von 0.011, leisten damit immerhin das Elffache an Varianzaufklärung gegenüber Modell A in Mathematik. In Modell B sind aufgrund der Fallzahlreduzierung nur noch die Universitäten Göttingen, Heidelberg und Münster enthalten. Mit erhöhtem Verhältnis der Hausarbeit zu den mündlichen Prüfungen, einer zunehmenden Gewichtung der Hausarbeit bei der Gesamtnotenberechnung und einem selbst gewähltem Thema (im Vergleich zu einem vorgegebenen Thema) verbessert sich die Abschlussnote. Mit erhöhter Dauer der Hausarbeit und unter Vergabe ganzer sowie halber Noten (im Vergleich zur 0.3/0.7 Differenzierung) verschlechtert sie sich. Den stärksten isolierten Effekt weist dabei die Gewichtung der Abschlussarbeit auf, sie ist neben der Zeitvariable aber auch die einzige Variable, die nicht als Dummy konzipiert ist, weshalb sich ein Vergleich der übrigen beta-Koeffizienten verbietet. Die erklärenden Variablen bieten dem Modell, das eine Gesamtgüte von r^2=0.11 besitzt, eine Verbesserung von 0.04.

Biologie Diplom

Für Biologie ergibt sich in Modell A eine ebenso geringe Erklärungskraft wie für Mathematik. Auch der r^2-Zugewinn, der durch den Einschluss der Nebenfachratio und des Verhältnisses Hausarbeit zu mündlichen Prüfungen entsteht, ist gleich (+0.001). Beide Variablen weisen den erwarteten Effekt auf, nur im Falle Letzterer ist dieser auch signifikant. Jedoch wird auch dieser Einfluss insignifikant, wird das Modell unter Senkung der Fallzahl um die übrigen potentiellen Einflussvariablen erweitert (Modell B ohne Karlsruhe und Berlin).

Unter Kontrolle des Hochschulstandorts und der Zeit ist jetzt nur noch für die beiden Dummy-Variablen ‚Thema der Abschlussarbeit selbst wählbar' (negatives Vorzeichen) und ‚Notensystem: Ganze Noten' (im Vergleich zur 0.3/0.7 Differenzierung, positives Vorzeichen) ein signifikanter Einfluss nachweisbar, die Modellgüte erhöht sich durch den Einschluss dieser beiden Faktoren um

0.006. Alle anderen Variablen sind insignifikant und tragen nicht zu einer Erhöhung des adjustierten r^2 von 0.03 bei.

Psychologie Diplom

Psychologie ist der Studiengang mit der höchsten Diversität formaler Prüfungsbedingungen und eingesetzter Prüfungsverfahren. Modell A enthält hier vier Einflussfaktoren, die allesamt einen signifikanten Effekt aufweisen. Mit steigender Anzahl Teilprüfungen und einem steigendem Anteil an Klausuren im Verhältnis zu mündlichen Prüfungen verschlechtert sich die Abschlussnote, mit abnehmender Anzahl mündlicher Prüfungen in Relation zur Hausarbeit und einem zunehmenden Nebenfachanteil verbessert sie sich. Letzteres sollte theoretisch eher zu einer Verschlechterung führen - warum ein erhöhter Nebenfachanteil notenverbessernd wirkt, muss an dieser Stelle erst einmal unklar bleiben. Da alle vier Variablen metrisch operationalisiert sind, ist auch ein Vergleich der standardisierten beta-Koeffizienten möglich. Dabei zeigt sich, dass die beiden notenverschlechternden Einflüsse einen deutlich stärkeren Effekt aufweisen als die beiden notenverbessernden. Das r^2 des Modells liegt mit 0.11 höher als es bei den A-Modellen in Mathematik und den beiden naturwissenschaftlichen Studiengängen der Fall ist, die vier erklärenden Variablen leisten eine Verbesserung um 0.015.
In Modell B gehen im Unterschied zu den vorherigen Studiengängen ebenfalls alle Hochschulen ein, die Fallzahl sinkt jedoch auch hier deutlich durch die Erweiterung der Liste aufgenommener Variablen. Die Anzahl an Teilprüfungen entfällt, da sie im verringerten Stichprobenumfang nun wieder perfekte Multikollinearität mit dem Verhältnis der Hausarbeit zur Anzahl mündlicher Prüfungen aufweist. Um die hohen VIF-Werte des Letzteren und der Gewichtung der Hausarbeit zu reduzieren, wird ein Interaktionsterm aus den beiden Variablen gebildet, der den Einfluss der Hausarbeit auf die Gesamtnote abbildet. Dieser weist den erwarteten negativen Effekt auf die Notenhöhe auf. Ebenfalls notenverbessernd wirken die Nutzung eines Notensystems mit ganzen oder halben Noten (im Vergleich zur 0.3/0.7 Differenzierung) und, wie bereits in Modell A, eine zunehmende Nebenfachratio. Als notenverschlechternder Einfluss verbleibt ein zunehmender Anteil Klausuren in Relation zur Anzahl mündlicher Prüfungen. Die übrigen Variablen (Dauer und Thema der Hausarbeit) sind nicht signifikant und tragen nicht zur Verbesserung der Modellgüte von r^2=0.08 bei, die im Gegensatz zu den vorher betrachteten Studiengängen geringer als in Modell A ist. Der Beitrag der erklärenden Variablen liegt mit 0.041 jedoch vergleichsweise hoch und beläuft sich auf die Hälfte der Erklärungskraft des Modells.

VWL Diplom

Auch in VWL weist ein Einflussfaktor nicht die erwartete Vorzeichenrichtung auf: Die Anzahl Klausuren im Verhältnis zur Anzahl mündlicher Prüfungen wirkt notenverbessernd statt -verschlechternd. Ebenfalls notenverbessernd wirkt eine verringerte Anzahl mündlicher Prüfungen gegenüber der

schriftlichen Abschlussarbeit, während ein steigender Nebenfachanteil die Abschlussnote erhöht. Dieses Modell (Modell A) weist bereits ein r^2 von 0.17 auf, was im Vergleich zu den anderen Studiengängen einen hohen Wert darstellt. Der Beitrag der erklärenden Variablen bleibt mit 0.007 jedoch sehr gering. In Modell B, das wie in Psychologie weiterhin alle Hochschulen unter Senkung der Fallzahl beinhaltet, liefern die erklärenden Variablen immerhin 0.021 des r^2 von 0.26. Der Nebenfachanteil und das Verhältnis Klausuren zu mündlichen Prüfungen bleiben unter stabilem Vorzeichen auch hier signifikant. Der Anteil mündlicher Prüfungen im Verhältnis zur schriftlichen Hausarbeit wird aufgrund hoher VIF-Werte wieder in einen Interaktionsterm mit der Gewichtung der Hausarbeit eingebracht werden. Auch hier zeigt sich wieder eine notenverbessernde Wirkung, genau wie bei der Dauer der Arbeit und der Bewertung nach einem 0.3/0.7 Notensystem (im Vergleich zur Nicht-Nutzung dieses Systems - aufgrund der hohen VIF-Werte können nicht die differenzierteren Notensystem-Dummys genutzt werden). Kann das Thema der Abschlussarbeit von den Kandidat*innen selbst gewählt werden, wirkt sich dies notenverschlechternd aus.

BWL Diplom

In BWL liegt die Erklärungskraft von Modell A mit r^2=0.12 in etwa so hoch wie in Chemie und Psychologie. Der positive Effekt der Nebenfachratio und der negative Effekt des Verhältnisses Hausarbeit zu mündliche Prüfungen leisten dabei allerdings einen wesentlich geringeren Anteil von 0.002 und sind damit ähnlich (wenig) erklärungskräftig wie in Mathematik und Biologie. In Modell B (ohne Tübingen) erhöht sich der Beitrag der erklärenden Variablen zur Gesamtgüte, die auf 0.11 sinkt, auf 0.006. Außer dem weiterhin positiven Nebenfachanteil zeigen sich ausschließlich notenverbessernde Wirkungen eines zunehmenden Gewichts der Abschlussarbeit bei der Gesamtnotenberechnung, einer längeren Bearbeitungsdauer und der Möglichkeit das Thema selbst zu wählen (im Vergleich zu einem vorgegebenen Thema). Die drei metrischen Variablen weisen etwa gleich starke Effekte auf (beta-Koeffizient). Das Verhältnis Hausarbeit zu mündlichen Prüfungen wird durch die Reduzierung der Fallzahl zu einer Konstanten und fällt damit aus dem Modell.

Tabelle 76: OLS Regression der Gesamtnote auf formale Prüfungsbedingungen und Prüfungsverfahren nach Studiengang

	Koeffizient	Standardfehler	beta	P>|t|	VIF[1]
Modell A	**Mathematik**				
Konstante	14.743	2.347		0.000	
Jahr	-0.007	0.001	-0.089	0.000	1.964
Göttingen[a]	0.013	0.030	0.007	0.649	1.496
Braunschweig[a]	-0.075	0.037	-0.029	0.042	1.406
Karlsruhe[a]	-0.104	0.033	-0.046	0.002	1.523
Berlin[a]	-0.207	0.033	-0.096	0.000	1.679
Tübingen[a]	-0.049	0.035	-0.019	0.161	1.287
Heidelberg[a]	-0.245	0.033	-0.111	0.000	1.523
Anteil NF vs. HF=1/2[b]	0.061	0.031	0.040	0.047	2.883
					n=6 834; r^2adj=0.02
Modell B	**Mathematik**				
Konstante	-1.405	3.881		0.717	
Jahr	0.002	0.002	0.024	0.366	2.912
Göttingen[a]	0.054	0.031	0.032	0.077	1.345
Braunschweig[a]	0.048	0.167	0.005	0.775	1.018
Tübingen[a]	-0.129	0.053	-0.044	0.016	1.343
Heidelberg[a]	-0.266	0.034	-0.147	0.000	1.445
Anteil NF vs. HF=1/2[b]	0.257	0.065	0.167	0.000	7.193
Gewicht der Arbeit	-1.338	0.704	-0.066	0.057	4.889
					n=3 983; r^2adj=0.02
Modell A	**Chemie**				
Konstante	17.154	0.983		0.000	
Jahr	-0.008	0.000	-0.132	0.000	1.074
Göttingen[a]	-0.350	0.020	-0.179	0.000	1.537
Braunschweig[a]	-0.488	0.035	-0.195	0.000	2.834
Karlsruhe[a]	0.010	0.033	0.005	0.753	3.809
Berlin[a]	-0.143	0.022	-0.064	0.000	1.372
Tübingen[a]	0.134	0.023	0.055	0.000	1.318
Heidelberg[a]	-0.184	0.020	-0.101	0.000	1.762
Anteil NF vs. HF=1/3[c]	-0.035	0.034	-0.013	0.294	2.244
Anteil Arbeit vs. mündlich=1/3[d]	-0.273	0.028	-0.185	0.000	5.118
					n=13 224; r^2adj=0.09
Modell B	**Chemie**				
Konstante	7.518	2.744		0.006	
Jahr	-0.002	0.001	-0.043	0.070	3.871
Göttingen[a]	-0.484	0.035	-0.302	0.000	3.204
Heidelberg[a]	-0.356	0.027	-0.257	0.000	2.646
Anteil NF vs. HF=1/3[c]	-0.008	0.038	-0.004	0.832	1.862
Anteil Arbeit vs. mündlich=1/3[d]	-0.119	0.033	-0.072	0.000	2.773
Gewicht der Arbeit	-1.134	0.320	-0.103	0.000	5.853
Dauer der Arbeit=9 Monate[f]	0.302	0.040	0.210	0.000	5.443
Thema selbst wählbar[g]	-0.122	0.041	-0.050	0.003	1.894
Notensystem[i]: 0.3/0.7 Differenzierung	-0.371	0.062	-0.270	0.000	13.965
					n=6 087; r^2adj=0.11

[1] Die einzelnen Prüfungsbedingungen sind oft sehr stark voneinander abhängig und erzeugen deshalb Kollinearitätsprobleme bei der Schätzung, wie die vereinzelt hohen VIF-Werte zeigen. Die grundsätzlichen Ergebnisse ändern sich bei Ausschluss einzelner Variablen zur Reduzierung dieser Werte jedoch nicht.

[a] Referenzkategorie: Münster

[b] Referenzkategorie: Anteil NF vs. HF=1/3

[c] Referenzkategorie: Anteil NF vs. HF=0

[d] Referenzkategorie: Anteil Arbeit vs. mündlich=1/4

[e] Referenzkategorie: Anteil Arbeit vs. mündlich=1/6

[f] Referenzkategorie: Dauer der Arbeit=6 Monate

[g] Referenzkategorie: Thema wird gestellt

[h] Referenzkategorie: Notensystem: 0.3/0.7 Differenzierung

[i] Referenzkategorie: Notensystem: keine 0.3/0.7 Differenzierung

noch Tabelle 76: OLS Regression der Gesamtnote auf formale Prüfungsbedingungen und Prüfungsverfahren nach Studiengang

	Koeffizient	Standardfehler	beta	P>|t|	VIF
Modell A	**Biologie**				
Konstante	20.635	1.675		0.000	
Jahr	-0.010	0.001	-0.113	0.000	1.197
Göttingen[a]	-0.027	0.020	-0.019	0.175	2.309
Braunschweig[a]	-0.042	0.024	-0.021	0.076	1.726
Karlsruhe[a]	-0.091	0.044	-0.032	0.037	2.748
Berlin[a]	-0.138	0.023	-0.095	0.000	3.090
Tübingen[a]	-0.204	0.032	-0.154	0.000	7.001
Heidelberg[a]	-0.217	0.036	-0.139	0.000	6.455
Anteil NF vs. HF	0.038	0.026	0.018	0.149	1.814
Anteil Arbeit vs. mündlich=1/3[d]	-0.115	0.031	-0.100	0.000	8.715
					n=11 651; r^2adj=0.02
Modell B	**Biologie**				
Konstante	3.818	5.298		0.471	
Jahr	-0.001	0.003	-0.010	0.672	3.209
Göttingen[a]	-0.077	0.025	-0.067	0.002	2.712
Braunschweig[a]	-0.086	0.072	-0.019	0.233	1.377
Tübingen[a]	-0.192	0.036	-0.086	0.000	1.464
Heidelberg[a]	-0.193	0.026	-0.155	0.000	2.415
Thema selbst wählbar[g]	-0.092	0.030	-0.071	0.002	3.040
Notensystem[h]: Ganze Noten	0.202	0.035	0.126	0.000	2.646
					n=5 375; r^2adj=0.03
Modell A	**Psychologie**				
Konstante	-2.932	1.724		0.089	
Jahr	0.002	0.001	0.026	0.026	1.906
Göttingen[a]	0.165	0.023	0.074	0.000	1.405
Braunschweig[a]	0.025	0.023	0.011	0.288	1.404
Berlin[a]	-0.225	0.016	-0.167	0.000	2.035
Tübingen[a]	0.144	0.017	0.090	0.000	1.583
Heidelberg[a]	0.153	0.017	0.095	0.000	1.601
Anzahl Teilprüfungen	0.090	0.011	0.129	0.000	3.666
Anteil NF vs. HF	-0.228	0.076	-0.029	0.002	1.243
Anteil Klausur(en) vs. mündlich	1.498	0.139	0.157	0.000	2.899
Anteil Arbeit vs. mündlich	-0.515	0.187	-0.036	0.006	2.386
					n=12 192; r^2adj=0.11
Modell B	**Psychologie**				
Konstante	1.400	2.803		0.617	
Jahr	0.000	0.001	0.004	0.851	4.217
Göttingen[a]	0.032	0.027	0.016	0.234	1.729
Braunschweig[a]	0.137	0.041	0.053	0.001	2.375
Berlin[a]	0.035	0.047	0.012	0.461	2.554
Tübingen[a]	0.310	0.032	0.214	0.000	4.554
Heidelberg[a]	0.310	0.031	0.213	0.000	4.193
Anteil NF vs. HF	-0.198	0.083	-0.028	0.017	1.292
Anteil Klausur(en) vs. mündlich	0.822	0.134	0.096	0.000	2.343
Anteil Arbeit vs. mündlich *Gewicht der Arbeit	-7.979	0.920	-0.187	0.000	4.403
Notensystem[h]: Ganze Noten	-0.380	0.034	-0.284	0.000	6.148
Notensystem[h]: Halbe Noten	-0.340	0.031	-0.162	0.000	2.124
					n=8 690; r^2adj=0.08

[a]Referenzkategorie: Münster
[b]Referenzkategorie: Anteil NF vs. HF=1/3
[c]Referenzkategorie: Anteil NF vs. HF=0
[d]Referenzkategorie: Anteil Arbeit vs. mündlich=1/4
[e]Referenzkategorie: Anteil Arbeit vs. mündlich=1/6
[f]Referenzkategorie: Dauer der Arbeit=6 Monate
[g]Referenzkategorie: Thema wird gestellt
[h]Referenzkategorie: Notensystem: 0.3/0.7 Differenzierung
[i]Referenzkategorie: Notensystem: keine 0.3/0.7 Differenzierung

noch Tabelle 76: OLS Regression der Gesamtnote auf formale Prüfungsbedingungen und Prüfungsverfahren nach Studiengang

	Koeffizient	Standardfehler	beta	P>|t|	VIF
Modell A			**VWL**		
Konstante	20.344	1.448		0.000	
Jahr	-0.009	0.001	-0.152	0.000	2.903
Göttingen[a]	0.095	0.021	0.040	0.000	1.315
Karlsruhe[a]	-0.925	0.047	-0.252	0.000	2.761
Berlin[a]	-0.575	0.016	-0.338	0.000	1.544
Tübingen[a]	-0.230	0.019	-0.114	0.000	1.462
Heidelberg[a]	-0.237	0.018	-0.120	0.000	1.453
Anteil NF vs. HF	0.164	0.053	0.055	0.002	5.236
Anteil Klausur(en) vs. mündlich	-0.604	0.084	-0.065	0.000	1.331
Anteil Arbeit vs. mündlich=1/5[e]	-0.104	0.026	-0.067	0.000	4.623
					n=13 766; r^2adj=0.17
Modell B			**VWL**		
Konstante	1.465	3.685		0.691	
Jahr	0.001	0.002	0.020	0.510	7.272
Göttingen[a]	0.001	0.034	0.001	0.971	3.271
Karlsruhe[a]	-0.922	0.086	-0.370	0.000	9.734
Berlin[a]	-0.566	0.033	-0.289	0.000	2.255
Tübingen[a]	0.177	0.065	0.057	0.007	3.667
Heidelberg[a]	-0.231	0.060	-0.095	0.000	4.995
Anteil NF vs. HF	0.158	0.053	0.070	0.003	4.400
Anteil Klausur(en) vs. mündlich	-0.824	0.088	-0.133	0.000	1.660
Anteil Arbeit vs. mündlich *Gewicht der Arbeit	-0.367	0.155	-0.051	0.018	3.801
Dauer der Arbeit	-0.047	0.016	-0.097	0.002	8.453
Thema selbst wählbar[g]	0.118	0.048	0.059	0.013	4.647
Notensystem[i]: 0.3/0.7 Differenzierung	-0.438	0.045	-0.295	0.000	7.440
					n=6 066; r^2adj=0.26
Modell A			**BWL**		
Konstante	18.188	0.891		0.000	
Jahr	-0.007	0.000	-0.119	0.000	1.665
Göttingen[a]	-0.025	0.010	-0.017	0.009	1.270
Berlin[a]	-0.514	0.011	-0.338	0.000	1.599
Tübingen[a]	-0.253	0.016	-0.100	0.000	1.200
Anteil NF vs. HF	0.066	0.014	0.034	0.000	1.531
Anteil Arbeit vs. mündlich	-2.600	0.533	-0.034	0.000	1.449
					n=26 242; r^2adj=0.12
Modell B			**BWL**		
Konstante	4.440	2.322		0.056	
Jahr	-0.001	0.001	-0.007	0.604	3.327
Göttingen[a]	-0.024	0.011	-0.018	0.034	1.475
Berlin[a]	-0.414	0.026	-0.240	0.000	4.343
Anteil NF vs. HF	0.125	0.023	0.041	0.000	1.143
Gewicht der Arbeit	-0.625	0.195	-0.042	0.001	3.360
Dauer der Arbeit	-0.042	0.008	-0.058	0.000	2.245
Thema selbst wählbar[g]	-0.198	0.026	-0.115	0.000	4.427
					n=17 577; r^2adj=0.11

[a]Referenzkategorie: Münster
[b]Referenzkategorie: Anteil NF vs. HF=1/3
[c]Referenzkategorie: Anteil NF vs. HF=0
[d]Referenzkategorie: Anteil Arbeit vs. mündlich=1/4
[e]Referenzkategorie: Anteil Arbeit vs. mündlich=1/6
[f]Referenzkategorie: Dauer der Arbeit=6 Monate
[g]Referenzkategorie: Thema wird gestellt
[h]Referenzkategorie: Notensystem: 0.3/0.7 Differenzierung
[i]Referenzkategorie: Notensystem: keine 0.3/0.7 Differenzierung

Die Daten zeigen also, dass die formalen Prüfungsbedingungen und die eingesetzten Prüfungsverfahren einen gewissen Einfluss auf die Notengebung besitzen. Studiengangübergreifend zeigt sich, dass die Abschlussnote besser wird, je geringer der Anteil mündlicher Prüfungen gegenüber der schriftlichen Hausarbeit ist, der Prüfungsform mit den im Durchschnitt besten Resultaten. Auch die (höhere) Gewichtung der Arbeit bei der Gesamtnotenberechnung wirkt generell notenverbessernd.

In welchem Ausmaß alleine die unterschiedliche Gewichtung der schriftlichen Arbeit (sowohl in der Gesamtnotenberechnung als auch in der Relation gegenüber der Anzahl der übrigen Teilprüfungen) Notendifferenzen begünstigt zeigt Tabelle 77. Spalte 2 gibt die Differenz der Mittelwerte der von 1960-1997 aggregierten Abschlussnoten für die in Spalte 1 angegeben Paarvergleiche an. In Spalte 3 ist der Wert angegeben, den diese Differenz annehmen würde, wenn die schriftliche Arbeit und die übrigen Teilprüfungen in den Studiengängen die gleiche Gewichtung erhalten würden. Um dies zu simulieren wurde für jeden Prüfling der Mittelwert aus den tatsächlich erzielten Teilprüfungsergebnissen (ohne die schriftliche Arbeit) berechnet. Anschließend wurde aus diesem (gewichtet mit dem Faktor 5 - der üblichen Anzahl Teilprüfungen ohne schriftliche Arbeit in BWL und VWL) und der (einfach gewichteten) Note der schriftlichen Arbeit eine fiktive Abschlussnote berechnet.

Es zeigt sich eine Verringerung der mittleren Notendifferenz um ca. eine Zehntelnote (Spalte 4) im Vergleich der mathematisch-naturwissenschaftlichen Studiengänge (die aufgrund der geringen Teilprüfungsanzahl von drei oder vier (üblicherweise mündlichen) Prüfungen die höchste faktische Gewichtung der schriftlichen Arbeit aufweisen) mit den wirtschaftswissenschaftlichen (fünf oder sechs Teilprüfungen). Noch deutlicher wird der Einfluss der Gewichtung, wenn nur die mathematisch-naturwissenschaftlichen Prüflinge, die drei Teilprüfungen absolviert haben, mit den wirtschaftswissenschaftlichen, die sechs Teilprüfungen absolviert haben, verglichen werden (Spalten 5 bis 7). Dabei ist zu berücksichtigen, dass die Differenz in den Prüfungsergebnissen zwischen mündlichen Prüfungen und Klausuren noch nicht herausgerechnet ist und sich bei einer Angleichung der Prüfungsverfahren - zumindest in der Simulation - eine weitere Verringerung ergeben würde.

Tabelle 77: Angleichung der Notenniveaus bei Angleichung der Gewichtung der schriftlichen Arbeit

	Differenz Gesamtnote (n=70 174)	Differenz Simulation_5TP (n=69 242)	Verringerung der Differenz	Differenz Gesamtnote (n= 23 913)	Differenz Simulation_6TP (n=23 389)	Verringerung der Differenz
VWL-Mathematik	1.02	0.92	0.10	1.16	1.02	0.14
VWL-Chemie	0.98	0.84	0.14	1.22	0.99	0.23
VWL-Biologie	1.28	1.18	0.10	1.46	1.34	0.12
BWL-Mathematik	1.13	1.06	0.07	1.47	1.37	0.10
BWL-Chemie	1.09	0.98	0.11	1.53	1.34	0.19
BWL-Biologie	1.39	1.32	0.07	1.77	1.69	0.08

Alle anderen überprüften Variablen weisen studiengangspezifische Auswirkungen auf die Notenhöhe auf (Tab.78). Allerdings können die durchgeführten Analysen lediglich als erste Annäherung verstanden werden - alle Studiengänge weisen eine geringe Modellgüte (am höchsten in VWL und Psychologie) und vor allem einen geringen Beitrag der erklärenden Variablen über die Hochschulzugehörigkeit

und den Zeittrend hinaus auf. Dies ist wohl vorwiegend der Natur der Daten geschuldet - dadurch, dass Änderungen in den Prüfungsordnungen einmalige Eingriffe sind, die in der Regel direkt mehrere potentiell prüfungsrelevante Bedingungen der Prüfung gleichzeitig ändern, entsteht ein immanentes Problem unerwünschter Drittvariableneffekte, da meist mehrere dichotome Variablen gleichzeitig ihre Ausprägung wechseln. Weiterführende Analysen könnten experimentelle oder quasi-experimentelle Designs verwenden, um dieses Problem zu lösen und isolierte Effekte besser darzustellen.

Weist der Querschnittsvergleich schon in Bezug auf die betrachteten Studiengänge Einschränkungen auf, so ist ein aussagekräftiger Vergleich von Hochschulen innerhalb der einzelnen Studiengänge anhand der vorliegenden Daten nicht möglich. Wie die Erweiterung der Variablenliste in den OLS-Modellen gezeigt hat, liegen Werte für die meisten potentiellen Einflussgrößen nicht für alle Hochschulen vor. Um den Einfluss der Änderungen formaler Prüfungsbedingungen im Zeitverlauf erfassen zu können, müssten sie, als einmalige Eingriffe in den Prüfungsprozess, die sie sind, als Interventionsanalysen modelliert werden. Dies ist aufgrund der enormen zeitlichen Überschneidung der Prüfungszeitpunkte und Abschlüsse von Studierenden mit verschiedenen, gleichzeitig gültigen Prüfungsordnungen mit den vorhandenen Daten jedoch nicht umsetzbar. Auch hier könnten (quasi-)experimentelle Designs konzipiert werden, um Abhilfe zu schaffen.

Tabelle 78: Einfluss anhand OLS überprüfter Faktoren auf die Höhe der Gesamtnote nach Studiengang

↗	Mathematik	Chemie	Biologie	Psychologie	VWL	BWL
Anzahl Teilprüfungen	--	--	--	↗	--	--
Anteil NF_HF	↗	--	--	↘	↗	↗
Anteil Klausur(en) vs. mündlich	--	--	--	↗	↘	--
Anteil Arbeit vs. mündlich	--	↘	↘	↘	↘	↘
Gewicht der Arbeit	--	↘	--	↘	↘	↘
Dauer der Arbeit	--	↗	--	--	↘	↘
Thema selbst wählbar	--	↘	↘	..	↗	↘
Notensystem[a] Ganze Noten	--	↗	↗	↘	--	--
Notensystem[a]: Halbe Noten	--	↗	--	↘	--	--
Notensystem[a]: Viertel Noten	--	--	--	--	--	--
Notensystem[b]: 0.3/0.7	--	--	--	--	↘	--

[a]Referenzkategorie: Notensystem: 0.3/0.7 Differenzierung
[b]Referenzkategorie: Notensystem: keine 0.3/0.7 Differenzierung

Prüfungsbelastung / Rahmenbedingungen für Lehre / Arbeitsmarktchancen der Prüflinge[105]

Im Abschnitt zu den potentiellen leistungskonformen Einflüssen auf die Notengebung wurde die Betreuungsrelation als Indikator für die Lehrqualität genutzt. Dafür lässt sich jedoch auch die Anzahl an Prüflingen verwenden: Sie entspricht abzüglich der Abbrecher*innen und Wechsler*innen der Anzahl an Studierenden der vorherigen Studienkohorte, also derjenigen, die ca. fünf Jahre zuvor ihr Studium begonnen haben. Sie bildet das Betreuungsverhältnis zwar nicht so genau ab, wie die exakte Relation zwischen Lehrenden und Studierenden, zeigt aber in den amtlichen Daten eine mittlere bis sehr hohe positive Korrelation mit der Betreuungsrelation auf Disziplinebene (Tab.79). Die Anzahl Prüflinge

[105] Teile des folgenden Abschnitts wurden bereits in Müller-Benedict/Gaens (2015) veröffentlicht.

erweist sich dabei als hochsignifikanter Prädiktor für die Höhe der Durchschnittsnote: Je größer die Prüfungsanzahl, umso schlechter ist die Note der Prüflinge im Mittel.

Tabelle 79: Korrelation zwischen Anzahl der Prüflinge und Betreuungsrelation nach Fach/Disziplin

Fach/Disziplin	Pearson's r
Wirtschaftswissenschaften	0.21
Biologie	0.28
Mathematik	0.33
Sozialwissenschaften	0.33
Psychologie	0.49
Maschinenbau, Verfahrenstechnik	0.63
Germanistik	0.64
Chemie	0.68
Rechtswissenschaften	0.82

Quelle: Hochschulpersonalstatistik, Hochschulprüfungsstatistik Berechnung durch Prof. Dr. Müller-Benedict

Tabelle 80: OLS-Regression der Durchschnittsnote auf die Anzahl Prüflinge

AV: Durchschnittsnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Anzahl Prüflinge	0.0002	8.97e-06	21.63	0.000
Jahr	-0.0139	0.0011	-12.75	0.000
Konstante	29.5017	2.1595	13.66	0.000
				n=599; r^2adj=0.49

Dass die Anzahl Prüflinge an dieser Stelle eingeführt wird und nicht im Abschnitt zu den leistungskonformen Einflüssen liegt daran, dass sie einen multiplen Indikator darstellt und sich auch als Indikator für leistungsexterne Einflüsse verwenden lässt. Äquivalent zu den Erwartungen der Lehrqualitätshypothese an die Daten (mit steigenden Prüfungszahlen steigen die Noten und umgekehrt) lassen sich auch Erwartungen zum leistungsexternen Einfluss der Prüfungsbelastung bzw. der Rahmenbedingungen für Lehre sowie zum Einfluss des Arbeitsmarkts formulieren: Möglich ist, dass die Lehrbedingungen, dargestellt anhand der Anzahl an Studierenden, nicht in dem Sinne auf die Lehrqualität wirken, dass sie die Noten leistungskonform senken oder erhöhen, sondern dass sie die Prüfenden dazu bewegen, gute Noten als Ausgleich für schlechte Lehr- und Lernbedingungen zu vergeben (Müller-Benedict/Tsarouha 2011). Sollte dies zutreffen, ist bei steigenden Prüfungszahlen mit sinkenden Noten zu rechnen[106]. Da die Lehrenden die Entwicklung der Lehrbedingungen direkt miterleben, müsste der Einfluss unmittelbar wirken. Im Gegensatz zum leistungskonformen Einfluss der Lehrqualität weist der vermutete Mechanismus eine Besonderheit auf: Mit steigenden Studierenden- bzw. Prüfungszahlen wird möglicherweise ein Absinken des Notenniveaus ausgelöst, umgekehrt dürfte mit sinkenden Prüfungszahlen jedoch keine Verschlechterung der Noten erfolgen - eine Verbesserung der Lehr- und Lernbedingungen wird Dozierende mit Sicherheit erfreuen, aber kaum eine*n von ihnen zu der Überlegung bringen, dass es gerecht wäre, die Noten zum Ausgleich für diese Verbesserung anzuheben.

[106] Auch Franz (2010) sieht sinkende Noten als Folge eines steigenden Prüfungsaufkommens. Er argumentiert, dass ein steigendes Prüfungsaufkommen Prüfende dazu bringen könnte, milder zu bewerten, um die Zeit, die sie eigentlich für Forschung und Verwaltung benötigen, nicht für Verhandlungen von Beschwerden über schlechte Noten nutzen zu müssen. Ein solcher Mechanismus könnte allerdings keine *kontinuierlichen* Verbesserungen generieren: Ab einem gewissen Grad an Verbesserung - bei dem die Beschwerdeanzahl auf das Minimum gesunken ist - gäbe es keinen Anlass mehr, die Noten weiter zu senken.

Der dritte Einfluss der Prüfungszahlen ergibt sich aus ihrer Zyklizität. Prüfungszahlen heute sind (leicht reduzierte) Erstsemesterzahlen eine Studiendauer zuvor. Eine Entscheidung für ein Studienfach wird neben einer intrinsischen Motivation auch von der Arbeitsmarktlage beeinflusst. Von ihr ist bekannt, dass sie bei vielen akademischen Berufen in langen Zyklen zwischen Überfüllung und Mangel schwankt (Titze 1990; Müller-Benedict 2005). Die Erstsemesterzahlen sind also Indikator für die Arbeitsmarktsituation: Solange sie steigen, herrscht Mangel, wenn sie wieder sinken, Überfüllung. Es gibt allerdings Fächer bzw. Studiengänge, die keine ausgeprägten Konjunkturen aufweisen. So ist der Ingenieur*innen- oder Lehrer*innenmangel vieldiskutiertes öffentliches Thema, aber wann konnte jemals in der Zeitung gelesen werden: ‚Eklatanter Soziolog*innenmangel befürchtet'?

Aber wie könnten die Berufsaussichten auf die Notengebung Einfluss nehmen? Hier gibt es zwei gegensätzliche Hypothesen. Die erste lautet: Wenn in einem Fach Arbeitsmarktüberfüllung herrscht, wird „milder" benotet. Daraus ergibt sich ein paralleler Verlauf der Erstsemester- und Notenzyklen: Solange jene sinken (Überfüllungsphase), sinken diese ebenfalls (werden besser), und umgekehrt. Die zweite Hypothese lautet: Wenn in einem Fach Arbeitsmarktüberfüllung herrscht, wird strenger selektiert (schlechter benotet). Damit erfolgt genau die gegensätzliche Bewegung: Solange die Erstsemesterzahlen sinken, steigen die Noten (werden schlechter). Für erstere Annahme gilt als Begründung, dass die Prüfenden Milde walten lassen wollen, um für die schlechten Aussichten zu trösten. (Hitpass/Trosien 1987:XI). Der zweiten Hypothese liegt die Vermutung zugrunde, dass wegen der Überfüllung ein strengeres Selektionsklima herrscht (Nath et al. 2004).

Die Prüfungszahlenkonjunktur folgt der Erstsemesterkonjunktur im Abstand einer Studiendauer. Die Zeitreihe der Prüfungszahlen, wenn sie um eine Studiendauerlänge in die Vergangenheit verschoben (gelagt) wird, zeigt so gerade die Arbeitsmarktkonjunkturen an: Sinkt die verschobene Zeitreihe, herrscht Überfüllung, wächst sie, herrscht Mangel.

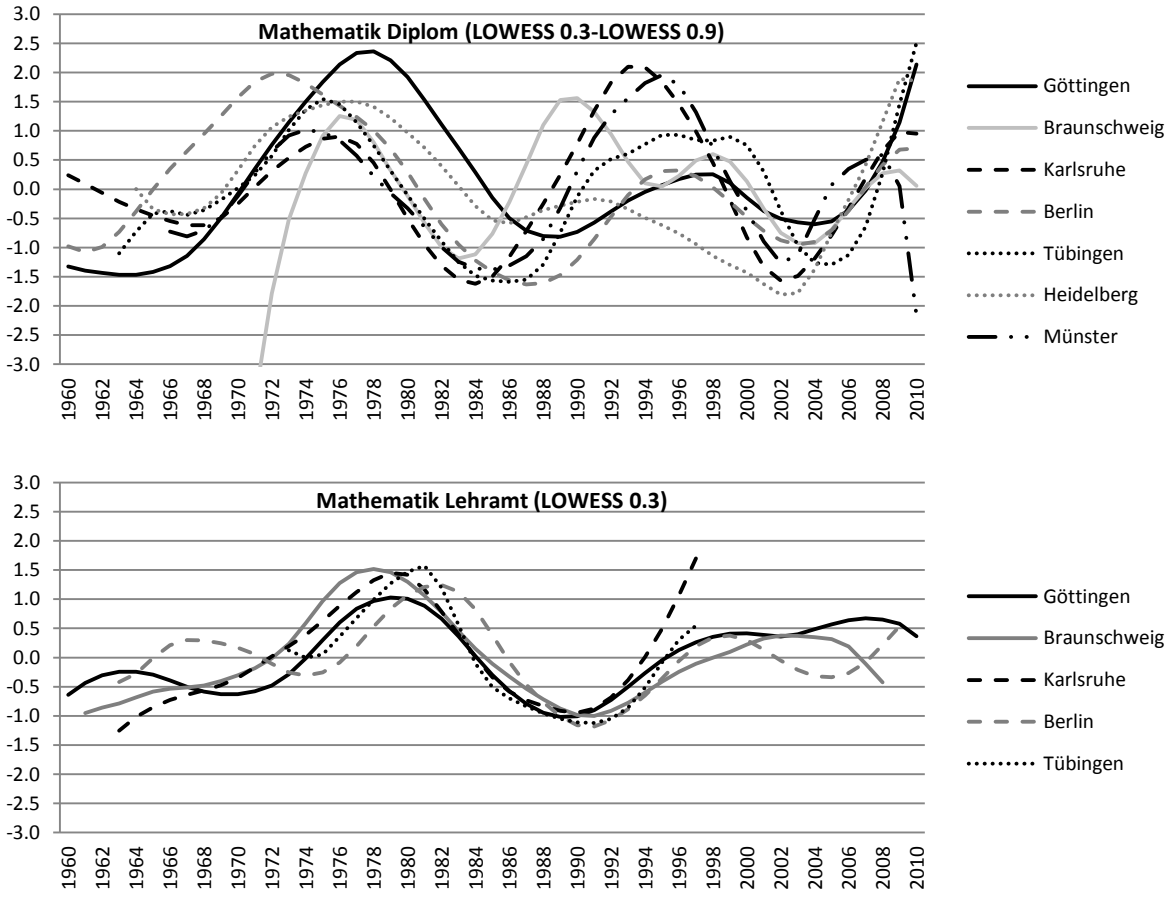
Welcher der beiden Einflüsse und welche der beiden Hypothesen sind eher mit den Daten vereinbar? Die Antwort hängt von der Lag-Struktur und dem Vorzeichen der Beziehung zwischen Prüfungszahlen und Noten ab. Wenn die Selektionshypothese gilt, müssten die gelagten Prüfungszahlen sich gegenläufig zu den Noten verändern (negatives Vorzeichen). Wenn die Mildehypothese gilt, müsste es genau umgekehrt sein: die gelagten Prüfungszahlen bewegen sich parallel zu den Noten (positives Vorzeichen). Wenn nicht der Arbeitsmarkt, sondern die Lehrbedingungen die Noten verändern, müssten die Noten sich genau synchron zu den Prüfungszahlen verhalten (kein Lag, positives Vorzeichen). Trifft die Ausgleichshypothese zu, ist eine gegenläufige Entwicklung zu erwarten (kein Lag, negatives Vorzeichen).

Um die Verhältnisse zu überprüfen, werden die Prüfungszahlen und die Abschlussnoten grafisch verglichen und es wird eine Regression durchgeführt. Da es sich bei den Daten um Zeitreihen mit autokorrelierten Residuen handelt, wird auch hier eine Prais-Winsten-Regression durchgeführt und zu-

sätzlich die Zeit als Variable berücksichtigt, um Notenverbesserung zu neutralisieren, wo nötig. Diese Regression wird mit Verschiebungen der Prüfungszahl in die Vergangenheit um 0 bis 5 Jahre durchgeführt, und das Lag mit der höchsten Anpassung (Signifikanz, R-Quadrat) gewählt[107]. Auf Studiengangebene ist die Analyse nur dort zulässig, wo auch einheitliche Fachkonjunkturen auftreten, sich die Prüfungszahlen also an allen Hochschulen synchron bewegen.

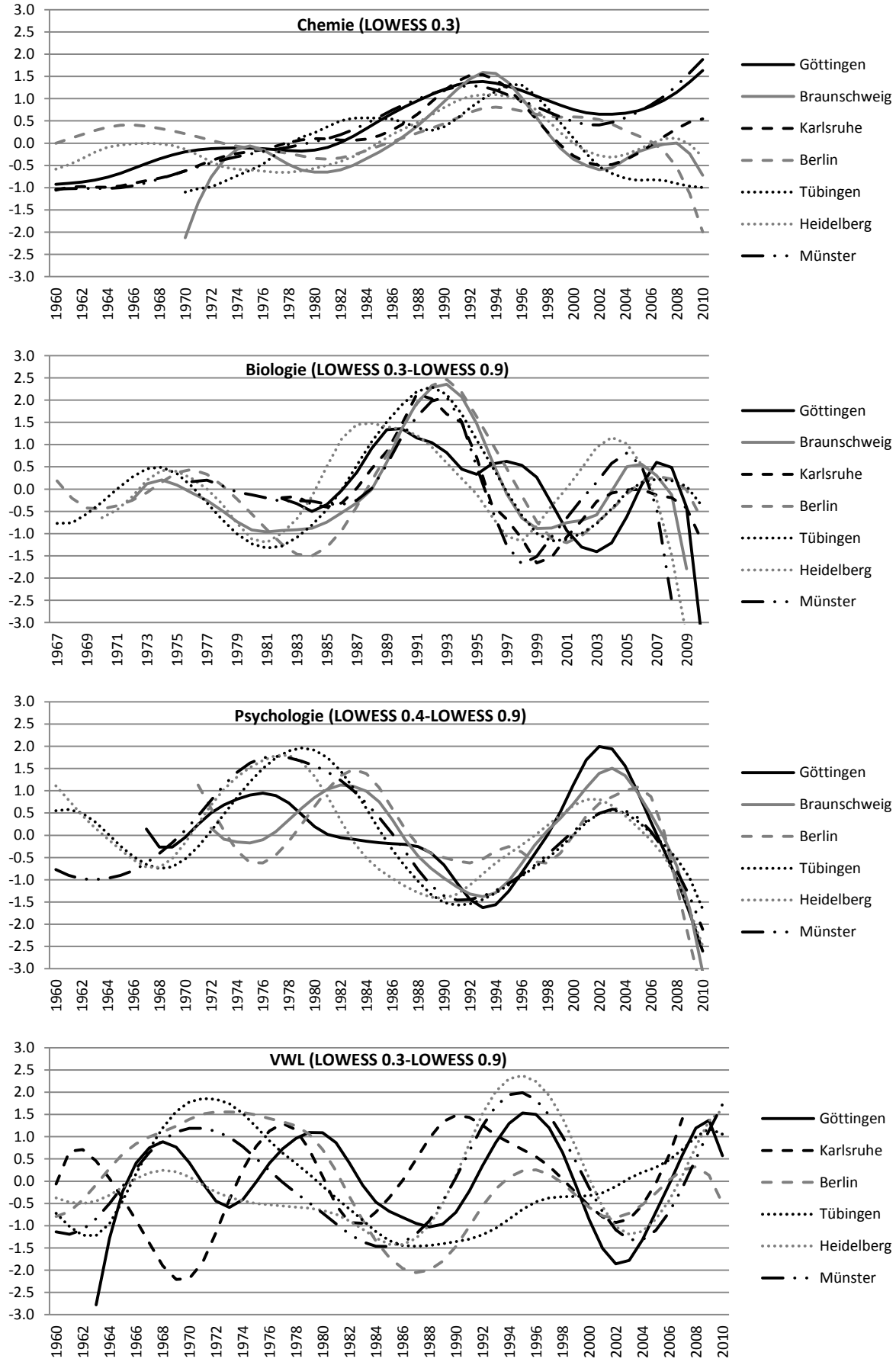
Im Folgenden ist die Entwicklung der Anzahl Prüflinge in den einzelnen Studiengängen, jeweils nach Hochschulen getrennt, dargestellt. Die Daten sind Z-standardisiert, um unabhängig vom absoluten Niveau der Zahlen einen Eindruck von ihrem Verlauf zu erhalten und somit die Schwankungen der Zeitreihen besser vergleichen zu können. Zudem sind die Zeitreihen gegebenenfalls trendbereinigt. Durch den Vergleich der Entwicklungen an den einzelnen Hochschulen zeigt sich, ob sich die Prüfungs- und damit die Studierendenzahlen hochschulübergreifend und damit studiengangspezifisch oder hochschulspezifisch entwickeln.

Abbildung 165: Entwicklung der Prüfungszahlen in den Studiengängen (z-standardisiert, ggf. trendbereinigt)

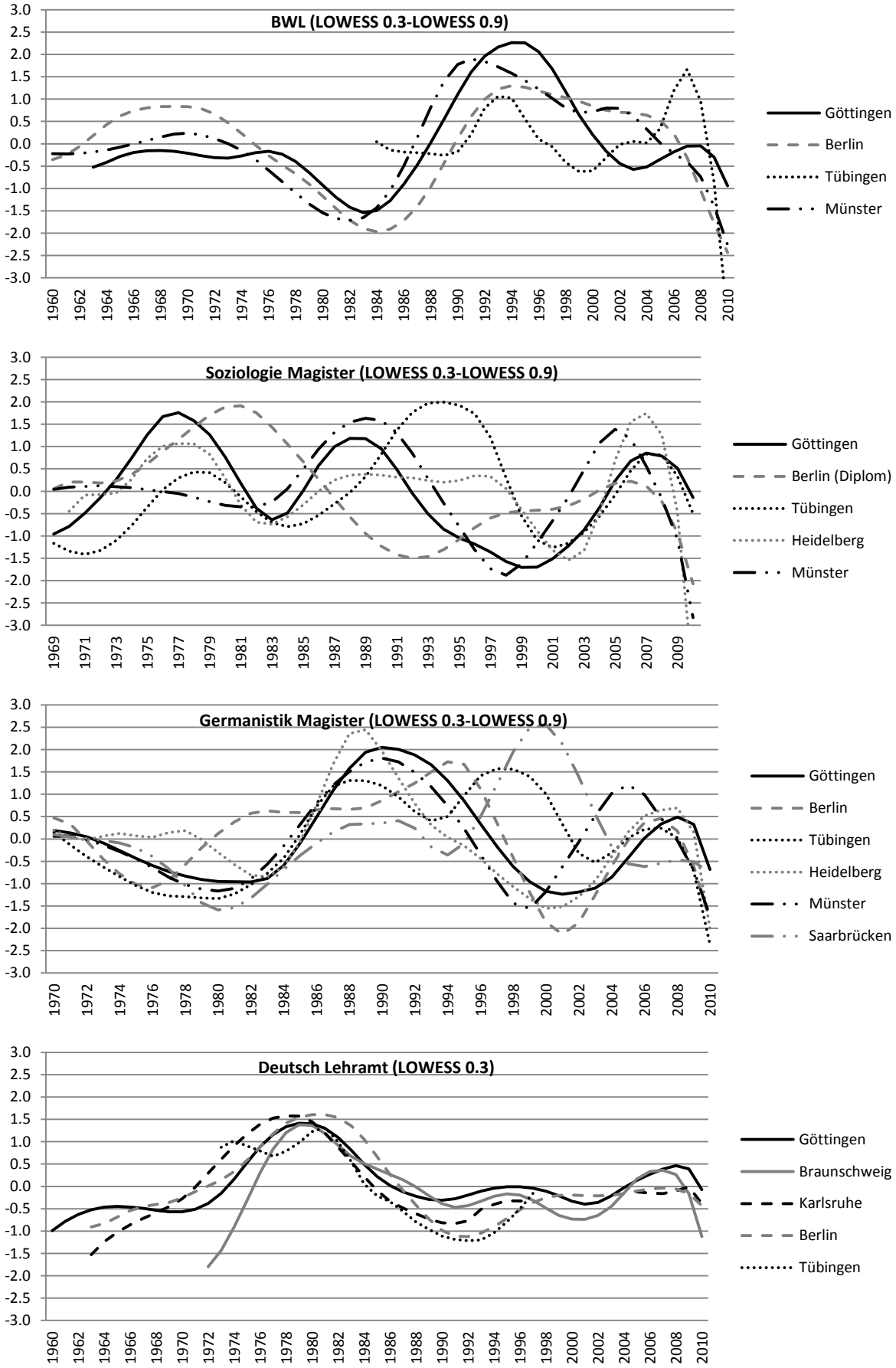


[107] Gerechnet wird im Folgenden immer mit den z-standardisierten, geglätteten Zeitreihen, die auch in den grafischen Abbildungen dargestellt sind. Die Feststellung signifikanter Abhängigkeiten wird durch die z-Standardisierung nicht berührt. Die Autokorrelation (ACF), die partielle Autokorrelation (PACF) und die unitroot (Dickey-Fuller-Test (DF)) aller Zeitreihen wurden geprüft und gegebenenfalls berücksichtigt. Deshalb wurde meist eine Prais-Winsten-Regression (PW-Regression) durchgeführt, um autokorrelierte Residuen zu neutralisieren, und der Durbin-Watson- Koeffizient (DW) geprüft.

noch Abbildung 165: Entwicklung der Prüfungszahlen in den Studiengängen (z-standardisiert, ggf. trendbereinigt)



noch Abbildung 165: Entwicklung der Prüfungszahlen in den Studiengängen (z-standardisiert, ggf. trendbereinigt



In allen Studiengängen zeigen sich zyklische Verläufe. Es zeigen sich hochschulübergreifend zeitlich parallele Entwicklungen der Prüfungszahlen in den Diplomstudiengängen Chemie, Biologie und BWL und in beiden Lehramtsstudiengängen. In Psychologie scheint der erste Zyklus in Braunschweig und Berlin gegenläufig zu den übrigen Standorten einzusetzen, ein Abgleich mit den Originaldaten zeigt jedoch, dass dieser Eindruck durch die Glättung der beiden später beginnenden Zeitreihen entsteht (siehe Anhang: Abb.A33-A34). In Mathematik Diplom gibt es eine leichte Verschiebung in der Braunschweiger Reihe, in der der zweite Zyklus etwas eher einsetzt als an den übrigen Hochschulen, sowie in Göttingen und Berlin, wo der erste Zyklus später bzw. früher seinen Peak erreicht. In VWL sind in den 1960er Jahren noch gegenläufige Bewegungen in den Karlsruher Zahlen vorhanden, die Reihe verläuft dann nach Angleichung der Verlaufsform in den 1970er Jahren auch in den 1980ern noch leicht versetzt. In den beiden Magisterstudiengängen verlaufen die Prüfungszahlen hingegen über den gesamten Zeitraum gegenläufig und damit eindeutig hochschulspezifisch.

Da die Lehrqualität unmittelbar an den Hochschulen Einfluss auf die Leistung der Studierenden ausübt, ist eine studiengangspezifische Analyse der Entwicklung nur dort sinnvoll, wo die Prüfungszahlen auch einheitlich verlaufen. Existiert ein genereller Einfluss der Lehrqualität an den einzelnen Hochschulen und wird dieser von den Prüfungszahlen abgebildet, müsste er sich bei einheitlichen Verlaufsmustern der Hochschulen innerhalb eines Studiengangs auch auf der aggregierten Studiengangebene nachweisen lassen. Nur in diesen Studiengängen kann dann auch eine studiengangspezifische Entwicklung der Noten im Zeitverlauf - zumindest theoretisch - mit einer studiengangspezifischen Entwicklung der Lehrqualität erklärt werden.

Die Ergebnisse zeigen, dass eine Entscheidung zwischen den Einflüssen und den Hypothesen je nach Studiengang bzw. Hochschule unterschiedlich ist. In den Grafiken (Abb.166-173) ist bereits gut zu erkennen, was sich dann in den Regressionsrechnungen (Tab.81-82[108]) bestätigt: In Mathematik Diplom und Lehramt, Deutsch Lehramt, Psychologie Diplom und VWL Diplom gilt die Selektionshypothese (negatives Vorzeichen zwischen zwei und fünf Jahren Verschiebung): Vom Minimum der verschobenen Prüfungszahlen beginnend (Mangelphase), verbessern sich (sinken) auch die Noten, und umgekehrt. Diese Situation kann damit in fünf von acht Studiengängen mit deutlichen Fachkonjunkturen bestätigt werden.

[108] Da mit dem LOWESS-Verfahren geglättete Zeitreihen eine hohe Autokorrelation aufweisen, ergeben sich bei einer Prais-Winsten-Regression LOWESS-geglätteter Zeitreihen nur schlechte (niedrige) DW-Werte, die Artefakte der Glättung sind. Deshalb wurden alle P-W-Regressionen in Tabelle 81 und 82 auch mit den Originaldaten überprüft: Die Rechnungen bestätigen die Ergebnisse und weisen dabei ausreichende D-W Werte auf (siehe Anhang: Tab.A22-A23).

In BWL, Chemie und Biologie zeigt sich für Lag0 die beste Passung (Tab.81), wobei in BWL und Chemie die Wirkung der Lehrbedingungen (kein Lag, positives Vorzeichen) sichtbar wird und die Daten lediglich in Biologie für das Vorliegen der Ausgleichshypothese (kein Lag, negatives Vorzeichen) sprechen. Dieses Ergebnis passt dazu, dass Biolog*innen trotz der sichtbar synchronen Prüfungszahlen eher einem diffusen Arbeitsmarkt ausgesetzt sind, der deshalb auch keinen Widerklang in den Prüfungsergebnissen finden sollte. Die Parallelität der Prüfungszahlenverläufe könnte durch die späte Einführung des Diplomstudiengangs Biologie an vielen Hochschulen zustande kommen, die einen gemeinsamen Aufschwung bis Anfang der 1990er Jahre mit sich bringt, der dann durch die Auslastung des zuvor kontinuierlich ausgebauten Studienplatzvolumens beendet wird. Die folgende Überfüllung wäre dann durch die NC-Begrenzung, nicht durch den Arbeitsmarkt indiziert.

Abbildung 166: Abschlussnoten/Prüflinge Mathematik Dip.

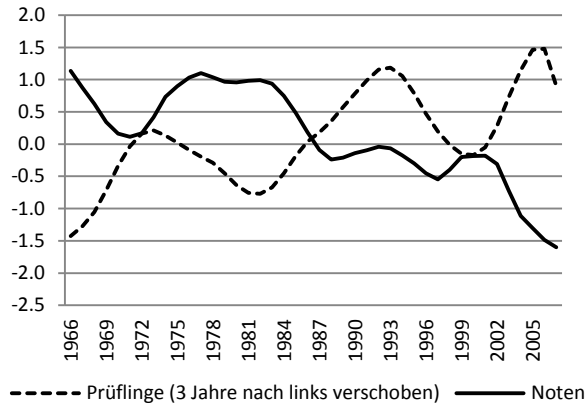


Abbildung 167: Abschlussnoten/Prüflinge Mathematik LA

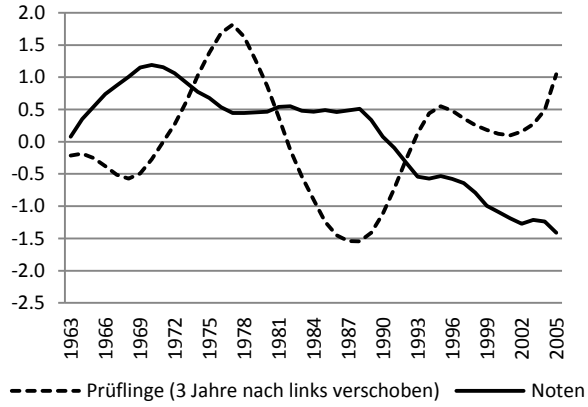


Abbildung 168: Abschlussnoten/Prüflinge Chemie Diplom

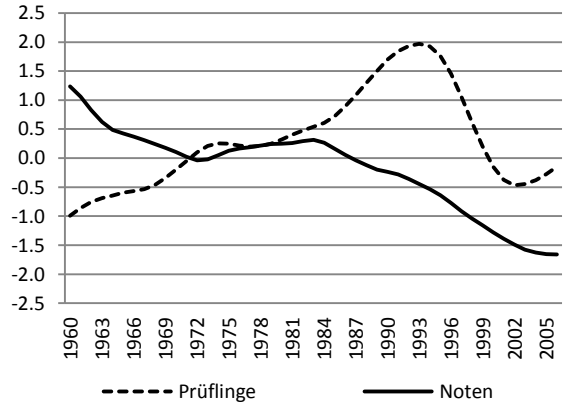


Abbildung 169: Abschlussnoten/Prüflinge Biologie Diplom

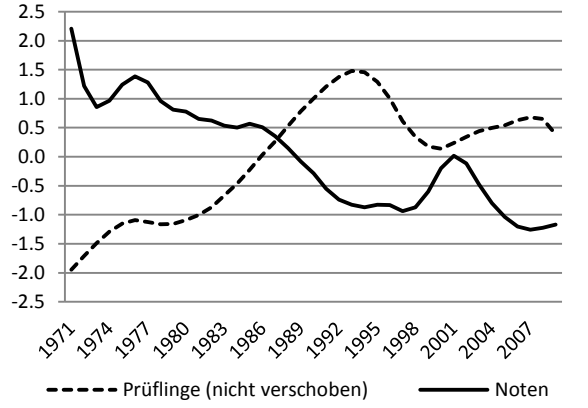


Abbildung 170: Abschlussnoten/Prüflinge Psychologie Dip.

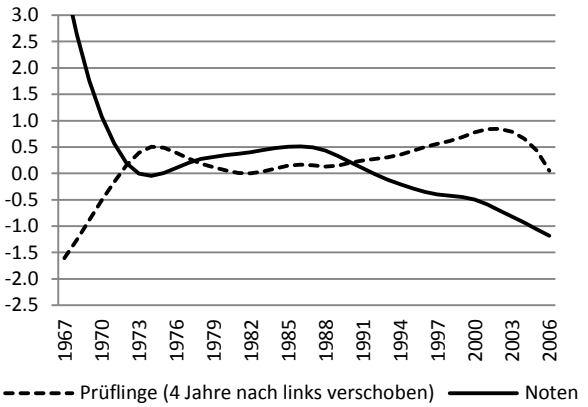


Abbildung 171: Abschlussnoten/Prüflinge VWL Diplom

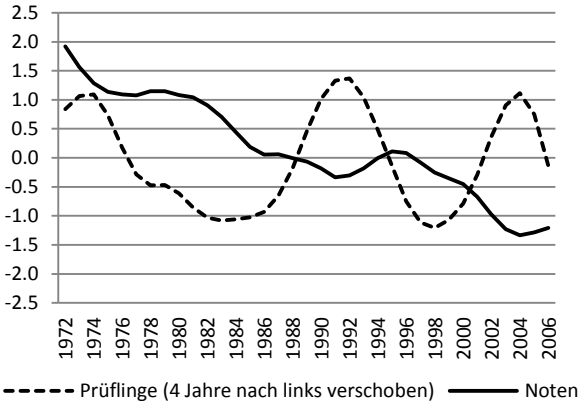


Abbildung 172: Abschlussnoten/Prüflinge BWL Diplom

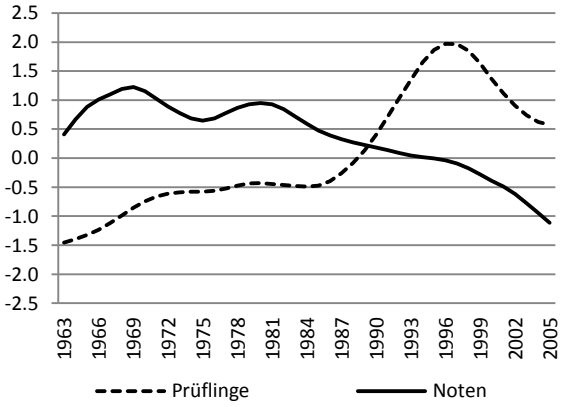


Abbildung 173: Abschlussnoten/Prüflinge Deutsch Lehramt

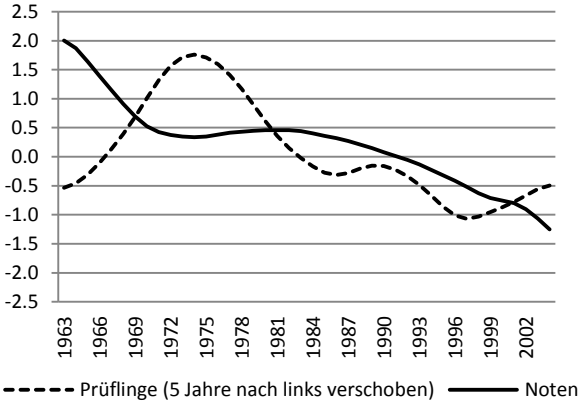


Tabelle 81: P-W-Regression der Abschlussnoten auf die Prüfungszahlen für Studiengänge mit Fachkonjunktur

	Koeffizient	Standardfehler	t-Statistik	P>|t
Note Mathematik Diplom				
Anzahl Prüflinge (Lead 3)	-0.328	0.108	-3.04	0.004
Jahr	-0.046	0.016	-2.78	0.008
Konstante	90.34	32.54	2.78	0.008
			D-W= 0.48; r^2_{adj}=0.33; n=42	
Note Mathematik Lehramt				
Anzahl Prüflinge (Lead 3)	-0.208	0.062	-3.37	0.002
Jahr	-0.030	0.015	-1.92	0.061
Konstante	58.27	30.53	1.91	0.063
			D-W= 0.66; r^2_{adj}=0.22; n=43	
Note Chemie Diplom				
Anzahl Prüflinge (Lead 0)	0.146	0.069	2.11	0.040
Jahr	-0.066	0.008	-7.80	0.000
Konstante	131.2	16.81	7.81	0.000
			D-W= 0.46; r^2_{adj}=0.49; n=51	
Note Biologie Diplom				
Anzahl Prüflinge (Lead 0)	-0.436	0.129	-3.39	0.002
Jahr	-0.048	0.011	-4.28	0.000
Konstante	96.04	22.43	4.28	0.000
			D-W= 1.18; r^2_{adj}=0.74; n=39	
Note Psychologie Diplom				
Anzahl Prüflinge (Lead 4)	-0.470	0.130	-3.61	0.001
Jahr	-0.060	0.020	-3.06	0.004
Konstante	120.0	39.15	3.06	0.004
			D-W= 0.57; r^2_{adj}=0.35; n=38	
Note VWL Diplom				
Anzahl Prüflinge (Lead 4)	-0.179	0.048	-3.70	0.001
Jahr	-0.093	0.011	-8.28	0.000
Konstante	185.87	22.43	8.29	0.000
			D-W= 0.72; r^2_{adj}=0.67; n=35	
Note BWL Diplom (bis 2000[a])				
Anzahl Prüflinge (Lead 0)	0.208	0.050	4.18	0.000
Jahr	-0.075	0.007	-10.3	0.000
Konstante	148.3	14.49	10.2	0.000
			D-W= 0.34; r^2_{adj}=0.56; n=48	
Note Deutsch Lehramt (bis 1998[a])				
Anzahl Prüflinge (Lead 5)	-0.291	0.046	-6.38	0.000
Jahr	-0.070	0.006	-12.47	0.000
Konstante	138.89	11.11	12.50	0.000
			D-W= 0.68; r^2_{adj}=0.49; n=31	

[a] Die Beschränkungen für BWL und Deutsch Lehramt wurden gewählt, weil im folgenden Zeitraum die Datengrundlage (FDZ-Daten) zu unsicher ist. Deutsch LA = Berlin, Göttingen, Karlsruhe.

Für Soziologie Magister und Germanistik Magister muss die Analyse wegen des offensichtlichen Fehlens einer Fachkonjunktur für einzelne Universitäten durchgeführt werden: In Germanistik zeigt sich in Göttingen und Berlin der Einfluss der Lehrbedingungen (positives Vorzeichen ohne Lag bzw. maximal bei Lag1), in Soziologie nur in Göttingen: Die Noten schwingen hier parallel zu den Prüfungszahlen (siehe auch Abb.174, 175 u. 178). Im Kontrast dazu zeigt sich in Germanistik in Heidelberg und Saarbrücken sowie in Soziologie in Heidelberg und Berlin ein gegenläufiger Zusammenhang ohne time-lag (größer 1). Hier findet sich also ein Beleg für die Ausgleichshypothese (siehe auch Abb.176-177 u. 179-180). Für die Mildehypothese wurde dagegen gar keine Bestätigung gefunden, für Tübingen und Münster findet sich in den Daten weder in Germanistik noch in Soziologie ein Zusammenhang.

Abbildung 174: Abschlussnoten/Prüflinge Germanistik Göttingen

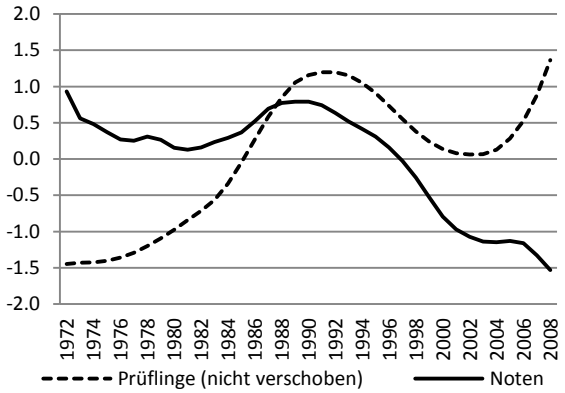


Abbildung 175: Abschlussnoten/Prüflinge Germanistik Berlin

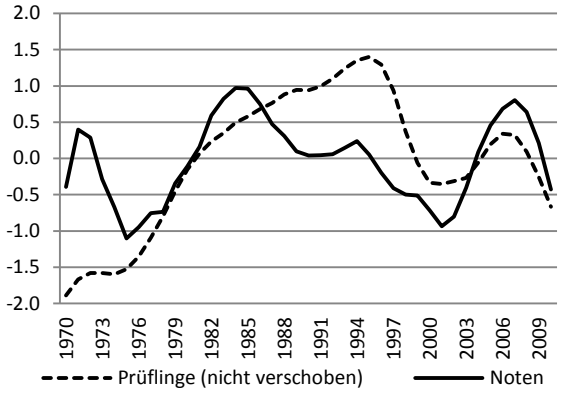


Abbildung 176: Abschlussnoten/Prüflinge Germanistik Heidelberg

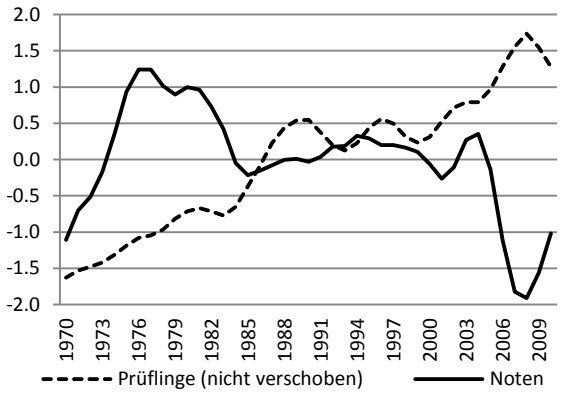


Abbildung 177: Abschlussnoten/Prüflinge Germanistik Saarbrücken

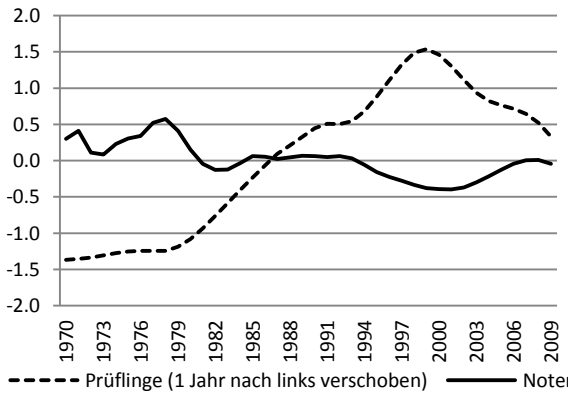


Abbildung 178: Abschlussnoten/Prüflinge Soziologie Göttingen

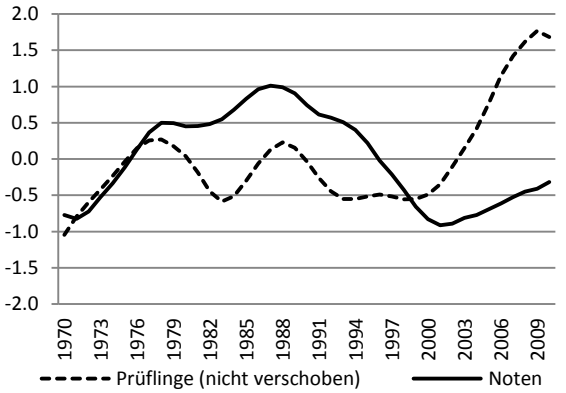


Abbildung 179: Abschlussnoten/Prüflinge Soziologie Berlin

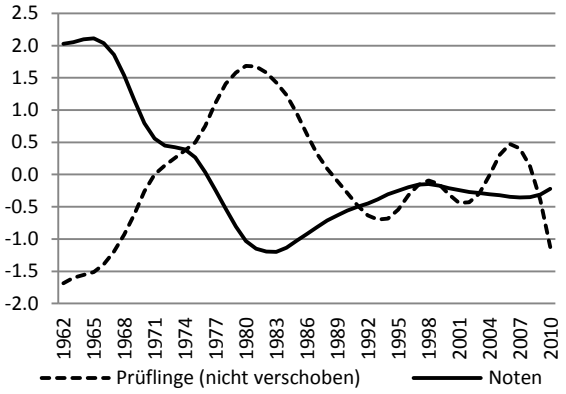


Abbildung 180: Abschlussnoten/Prüflinge Soziologie Heidelberg

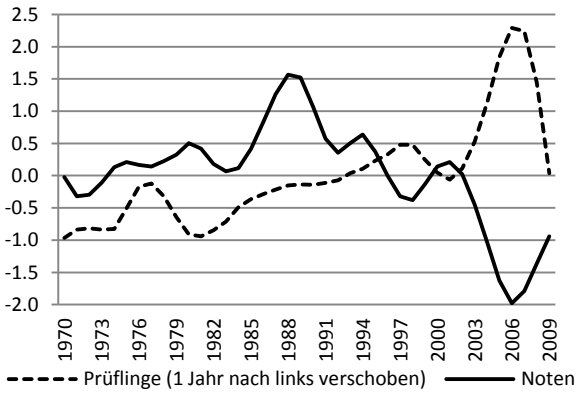


Tabelle 82: P-W-Regression der Abschlussnoten auf die Prüfungszahlen für Studiengänge ohne Fachkonjunktur

	Koeffizient	Standardfehler	t-Statistik	P>|t
Note Germanistik Magister Göttingen				
Anzahl Prüflinge (Lead 0)	0.591	0.127	4.65	0.000
Jahr	-0.101	0.015	-6.70	0.000
Konstante	200.9	29.94	6.71	0.000
			D-W= 0.44; r^2_{adj}=0.42; n=42	
Note Germanistik Magister Berlin				
Anzahl Prüflinge (Lead 0)	0.731	0.177	4.13	0.000
Jahr	-0.023	0.023	-1.01	0.321
Konstante	46.05	45.63	1.01	0.319
			D-W= 0.94; r^2_{adj}=0.28; n=41	
Note Germanistik Magister Heidelberg				
Anzahl Prüflinge (Lead 0)	-0.775	0.334	-2.32	0.026
Jahr	0.052	0.038	1.37	0.177
Konstante	-104.21	75.47	-1.38	0.175
			D-W= 0.68; r^2_{adj}=0.09; n=41	
Note Germanistik Magister Saarbrücken				
Anzahl Prüflinge (Lead 1)	-0.235	0.088	-2.67	0.011
Jahr	0.002	0.007	0.30	0.768
Konstante	-4.124	13.95	-0.30	0.769
			D-W= 1.16; r^2_{adj}=0.28; n=40	
Note Soziologie Magister Göttingen				
Anzahl Prüflinge (Lead 0)	0.363	0.112	3.23	0.003
Jahr	-0.014	0.019	-0.71	0.481
Konstante	26.43	37.93	0.70	0.490
			D-W= 0.28; r^2_{adj}=0.18; n=41	
Note Soziologie Magister Berlin				
Anzahl Prüflinge (Lead 0)	-0.389	0.063	-6.21	0.000
Jahr	-0.042	0.012	-3.47	0.001
Konstante	83.90	24.09	3.48	0.001
			D-W= 0.28; r^2_{adj}=0.49; n=49	
Note Soziologie Magister Heidelberg				
Anzahl Prüflinge (Lead 1)	-0.508	0.113	-4.50	0.000
Jahr	-0.009	0.024	-0.38	0.708
Konstante	18.10	48.71	0.37	0.712
			D-W= 0.63; r^2_{adj}=0.33; n=40	

Besonderheiten der Arbeitsmärkte und der Professionen, die für die starken Unterschiede zwischen Fächern bzw. Studiengängen verantwortlich sein könnten, ließen sich zahlreiche und widersprüchliche aufzählen (Whitley 1984; Becher 1989). Interessant ist die oben erwähnte Zweiteilung in einerseits ihrem spezifischen Arbeitsmarkt „hart“ unterworfene und andererseits eher lose mit einem diffusen Arbeitsmarkt gekoppelte Berufe. Dass Studierende der Studiengänge dieser beiden Kategorien auf die Arbeitsmarktlagen unterschiedlich reagieren, haben Reisz/Stock (2013) nachgewiesen. Lehramtsstudiengänge gehören eher zur ersten, Germanistik und Soziologie als Magisterstudiengänge eher zur zweiten Gruppe.

Wenn auch noch unklar ist, wie die Arbeitsmarktlage oder ein überfüllter Studiengang auf die konkrete Benotungspraxis der Prüfenden einwirken, wurde der Zusammenhang zwischen Prüfungszahlen und Prüfungserfolg auch schon früher sowohl auf der Grundlage historischer Daten (Titze 1990; Müller-Benedict 2005) als auch bei der ersten langfristigen Analyse der bundesdeutschen Hochschul-

noten durch Hitpass/Trosien (1987) festgestellt. Mit diesen Analysen wird der Zusammenhang wieder bestätigt, allerdings zeigt er sich studiengangspezifisch. Im Falle der Arbeitsmarktkonjunktur ist der Einfluss außerhalb der Hochschulen selbst angesiedelt und wirkt trotzdem auf die Prüfungen.

Langfristig lässt sich also ein Zusammenhang zwischen Prüfungs- bzw. Erstsemesterzahlen und den Abschlussnoten nachweisen. Damit lassen sich die zyklischen Bewegungen der Noten erklären, die in allen Zeitreihen auftreten. Darüber hinaus kann jedoch ein Phänomen nachgewiesen werden, das mit der nachgewiesenen langfristigen Verbesserung der Noten in den meisten Studiengängen in Zusammenhang steht: Eine unterschiedliche Elastizität der Examensnoten nach oben und unten in ihrem Verlauf auf der Notenskala. Sie lässt sich bereits in der grafischen Analyse erkennen: In allen Abbildungen mit gegenläufigem Zusammenhang ist sichtbar, dass die Noten beim Ansteigen der (gegebenenfalls gelagten) Prüfungszahlen in der Regel sinken, beim Absinken der Prüfungszahlen jedoch nicht in gleichem Maße oder auch überhaupt nicht sichtbar steigen.

Dieses Phänomen kann durch eine besondere Berechnung entschleiert werden, die die Reaktion auf die Veränderung der Prüfungszahlen in zwei Komponenten zerlegt: Den Einfluss von Erhöhung und den Einfluss von Verringerung. Weil es um eine Beziehung zwischen zwei Veränderungen geht, wird mit den ersten Differenzen der Zeitreihen gerechnet. Die Veränderung der nach den Ergebnissen aus dem vorigen Abschnitt gelagten Prüfungszahl wurde in zwei Variablen, positives (steigende Zahlen) bzw. negatives (fallende Zahlen) Wachstum, aufgespalten[109] und eine Regression der Veränderung der Noten auf beide berechnet. Weil mit den ersten Differenzen der Daten gerechnet wird, ist hier keine Prais-Winsten-Regression nötig, es werden OLS-Regressionen berechnet. Die Ergebnisse zeigen die Abbildungen 181 bis 190 sowie die Tabellen 83 und 84.

[109] Für die Phasen des jeweils anderen Wachstums wurden die Variablen auf 0 gesetzt. Simulationsrechnungen zeigen, dass die mit diesem Verfahren gewonnenen zwei Koeffizienten gemittelt genau den Wert des Koeffizienten der nicht aufgespaltenen Variable ergeben.

Abbildung 181: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Mathematik Dip.

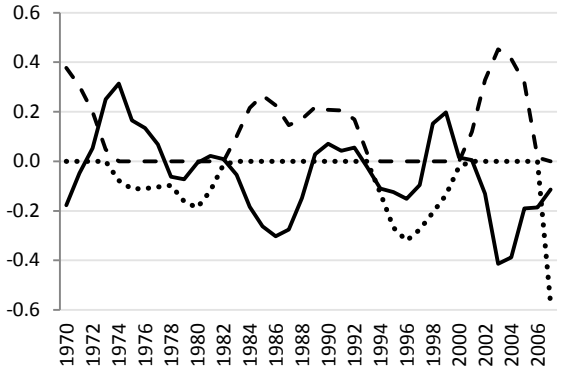


Abbildung 182: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Mathematik LA

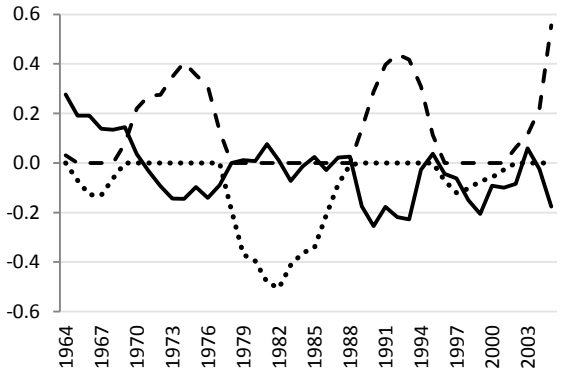


Abbildung 183: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Chemie Diplom

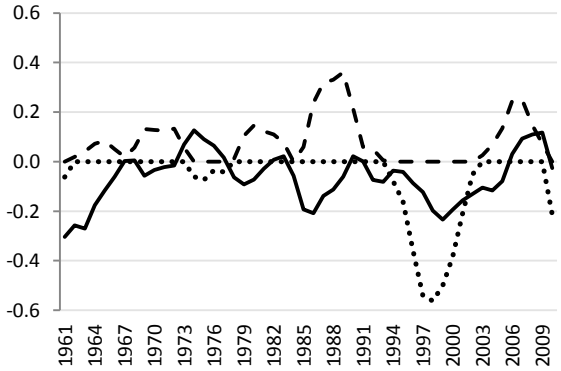


Abbildung 184: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Biologie Diplom

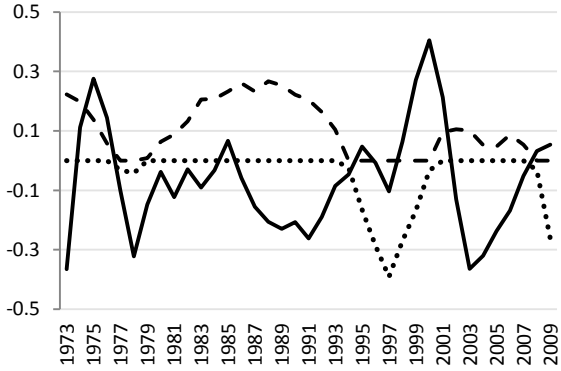


Abbildung 185: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Psychologie Dip.

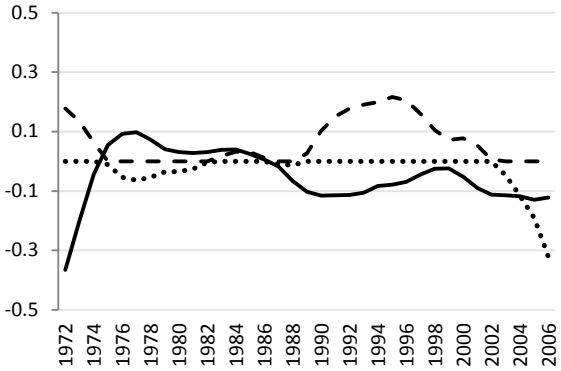


Abbildung 186: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) VWL Diplom

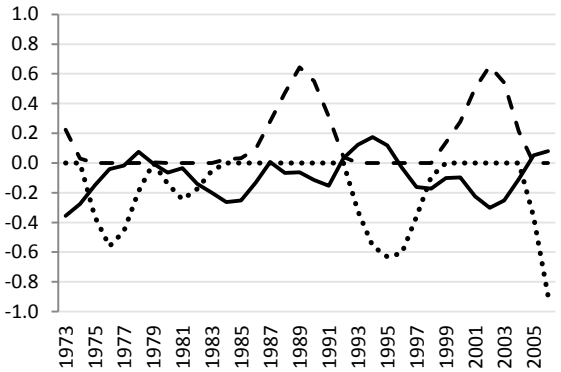


Abbildung 187: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) BWL Diplom

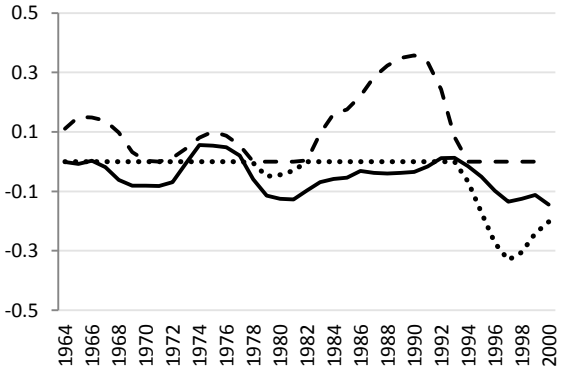


Abbildung 188: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Deutsch Lehramt

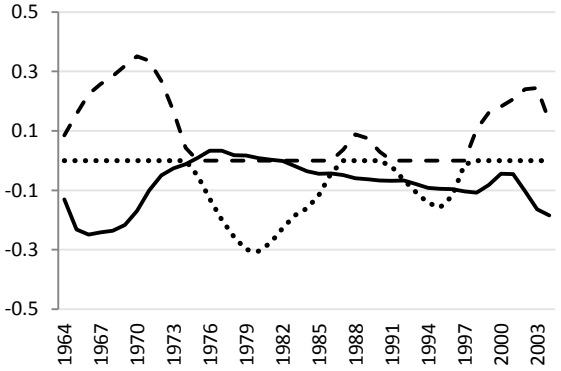


Tabelle 83: OLS-Regression der Abschlussnoten auf das Wachstum der Prüfungszahlen für Studiengänge mit Fachkonjunktur

	Koeffizient	Standardfehler	t-Statistik	P>|t
D.Note Mathematik Diplom				
positives Wachstum	-0.628	0.115	-5.46	0.000
negatives Wachstum	0.101	0.161	0.63	0.534
				r2=0.44; n=41
D.Note Mathematik Lehramt				
positives Wachstum	-0.377	0.057	-6.60	0.000
negatives Wachstum	0.023	0.067	0.35	0.731
				r2=0.56; n=36
D.Note Chemie Diplom				
positives Wachstum	-0.262	0.118	-2.23	0.030
negatives Wachstum	0.358	0.092	3.90	0.000
				r2=0.30; n=50
D.Note Biologie Diplom				
positives Wachstum	-0.851	0.251	-3.39	0.002
negatives Wachstum	-0.088	0.333	-0.26	0.793
				r2=0.24; n=38
D.Note Psychologie Diplom				
positives Wachstum	-1.321	0.161	-8.20	0.000
negatives Wachstum	0.231	0.232	1.00	0.326
				r2=0.65; n=39
D.Note VWL Diplom				
positives Wachstum	-0.352	0.088	-3.99	0.000
negatives Wachstum	-0.027	0.764	-0.35	0.727
				r2=0.33; n=34
D.Note BWL Diplom				
positives Wachstum	-0.097	0.058	-1.68	0.101
negatives Wachstum	0.460	0.082	5.60	0.000
				r2=0.49; n=37
D.Note Deutsch Lehramt				
positives Wachstum	-0.461	0.308	-1.49	0.147
negatives Wachstum	0.072	0.309	0.23	0.817
				r2=0.08; n=29

Tabelle 84: OLS-Regression der Abschlussnoten auf das Wachstum der Prüfungszahlen für Studiengänge ohne Fachkonjunktur

	Koeffizient	Standardfehler	t-Statistik	P>|t
D.Note Germanistik Magister Göttingen				
positives Wachstum	0.098	0.173	0.57	0.575
negatives Wachstum	0.997	0.328	3.03	0.004
				r^2=0.19; n=41
D.Note Germanistik Magister Heidelberg				
positives Wachstum	-0.705	0.326	-2.16	0.038
negatives Wachstum	-0.521	0.596	-0.87	0.388
				r^2=0.14; n=36
D.Note Germanistik Magister Saarbrücken				
positives Wachstum	-0.265	0.146	-1.81	0.078
negatives Wachstum	-0.169	0.236	-0.72	0.477
				r^2=0.09; n=39
D.Note Soziologie Diplom Berlin				
positives Wachstum	-0.680	0.086	-7.87	0.000
negatives Wachstum	-0.173	0.076	-2.28	0.027
				r^2=0.59; n=48
D.Note Soziologie Magister Heidelberg				
positives Wachstum	-0.721	0.179	-4.02	0.000
negatives Wachstum	-0.394	0.136	-2.91	0.006
				r^2=0.40; n=39

Abbildung 189: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Germanistik Gött.

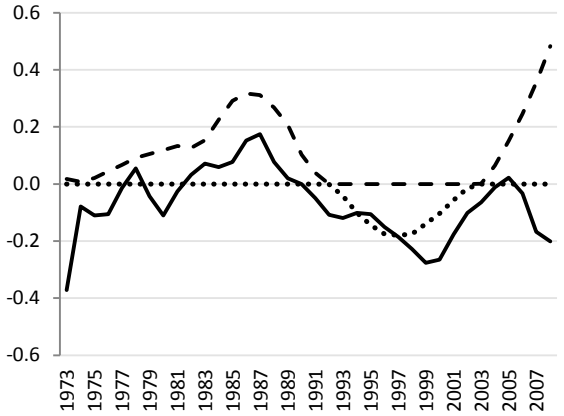


Abbildung 190: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Germanistik H'berg.

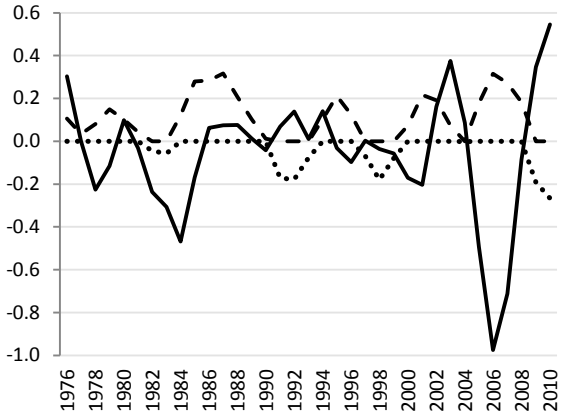


Abbildung 191: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Germanistik Saarbr.

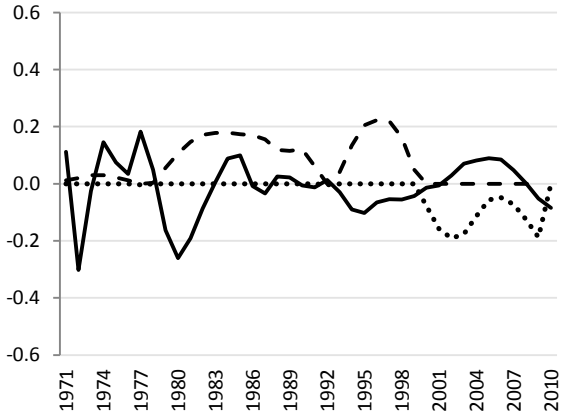


Abbildung 192: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Soziologie Berlin

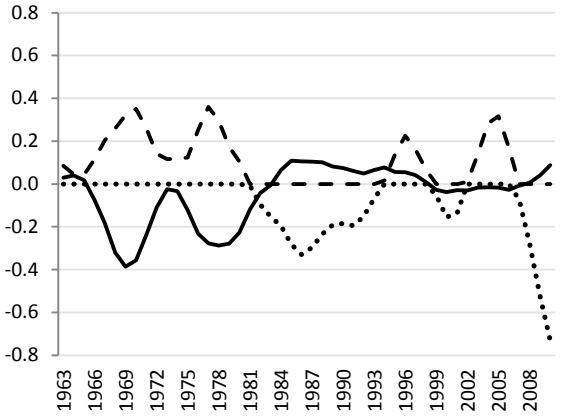
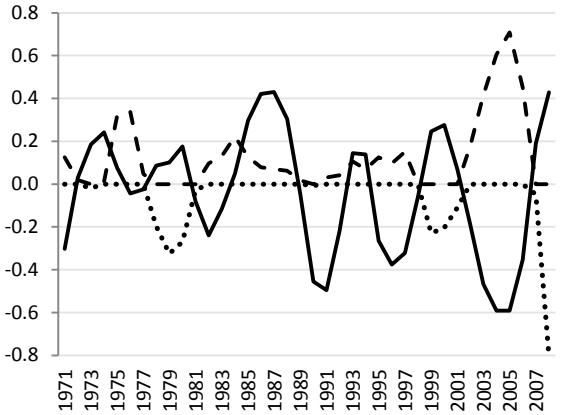


Abbildung 193: Wachstum Noten (durchgehend) und Prüfungszahlen (gestrichelt – positiv, gepunktet – negativ) Soziologie H'delberg



In allen Studiengängen, für die die Selektionshypothese bestätigt werden konnte, werden in der Mangelphase (wenn die - verschobenen - Prüfungszahlen wachsen) die Noten besser (geringer) (in Mathematik Lehramt zum Beispiel um den Betrag 0.377 pro Standardabweichung), in der Überfüllungsphase dagegen bleiben sie gleich (0.023 in Mathematik Lehramt ist nicht signifikant von 0 verschieden). Das bedeutet, die Noten sind unterschiedlich elastisch in Bezug auf die Phasen: Auf dieselbe Veränderung der Prüfungsanzahl verändern sich die Noten nur in Richtung „besser".
Hervorzuheben ist, dass sich für die Studiengänge BWL Diplom und Chemie Diplom, in denen die Lehrbedingungsthese bestätigt wurde und in denen daher der positive Zusammenhang zwischen Schrumpfung (negativem Wachstum) und Notenverbesserung überwiegt, dieselbe unterschiedliche Elastizität ergibt, obwohl die Noten anders mit der Prüfungsanzahl gekoppelt sind. Auch für Biologie Diplom, wo die Noten bei einer Verschlechterung der Lehrbedingungen stärker besser werden (um 0.851 Standardabweichungen, p=0.002), als sie sich bei deren Verbesserung verschlechtern (der Koeffizient von -0.088 ist nicht signifikant), bestätigt sich dieses Ergebnis.
Und auch auf Hochschulebene wirkt die unterschiedliche Elastizität: In Germanistik Magister an der Universität Göttingen (einer von zwei Hochschulen in Germanistik mit langfristiger Verbesserung - für Tübingen konnte kein Zusammenhang zwischen Noten und Prüfungszahlen nachgewiesen werden) stagnieren die Noten, wenn die Prüfungszahlen steigen (der Koeffizient von 0.098 ist ebenfalls nicht signifikant), in der Überfüllung jedoch verbessen sie sich (sinken) um 0.997 Standardabweichungen. Also ist auch bei dieser Art des Zusammenhangs der Prüfungszahlen mit den Noten die Reaktion in der Phase der Verbesserung der Noten stärker als in der Phase der Verschlechterung. Demnach besteht auch hier die ständige Tendenz zur Verbesserung. Auch in Heidelberg und Saarbrücken, wo die Noten ab Ende der 1970er Jahre in zyklischen Bewegungen besser werden, zeigt sich die unterschiedliche Elastizität der Noten gemäß der Ausgleichshypothese (stärkere Verbesserung der Noten bei Verschlechterung der Lehrbedingungen als Verschlechterung bei Verbesserung dieser).
Dadurch wird das Niveau nach jedem Zyklus ein wenig besser und über die ganze Zeitspanne hinweg ergibt sich eine Inflation der Noten. Für Soziologie Magister bestätigt sich das Ergebnis in Heidelberg und damit am einzigen Hochschulstandort, der eine langfristige Verbesserung der Noten in diesem Studiengang aufweist, für Diplom an der FU Berlin, wo sich das Notenniveau zu Beginn der Reihe erst einmal über längere Zeit senkt.

Dieses Ergebnis könnte eine mögliche Ursache für den langfristigen Trend zur grade inflation sein. In den Verbesserungsphasen verbessern sich die Noten stärker, als sie sich in den Verschlechterungsphasen verschlechtern. Dies ist möglicherweise auch die Ursache dafür, dass das einmal nach einer Mangelphase um ca. 1980 erreichte sehr gute Notenniveau in Biologie und Psychologie sich seitdem

konstant hält und nicht wieder hebt. Um es mit Zirkel (1999:255) zu sagen: „the basic problem is that high grades are simply easier".

Dass unterschiedliche Arbeitsmarktchancen bzw. Lehrbedingungen und die daraus resultierenden Selektionsklimata - nun wieder aus einer Querschnittsperspektive betrachtet - zu Differenzen im Notenniveau zwischen Studiengängen führen, ist (mit Ausnahme der leistungskonformen Lehrbedingungsthese) hingegen keine plausible Annahme. Denn in allen Studiengängen mit einheitlicher Fachkonjunktur entwickelt sich das Notenniveau ja in Abhängigkeit der studiengang- bzw. fachspezifischen Arbeitslosigkeit, also in intrafachlichen intertemporären Maßstäben und nicht orientiert an der tatsächlichen Quotenhöhe. Es könnten deshalb höchstens zeitliche Koinzidenzen der Phasen von Überfüllung und Mangel die Querschnittsunterschiede im Notenniveau zwischen Studiengängen beeinflussen, hier läge dann allerdings kein systematischer Einfluss vor. Für die nicht arbeitsmarktabhängigen Karrieren Biologie, Germanistik Magister und Soziologie Magister ist aufgrund der Arbeitsmarktsituation erst recht nicht mit Erklärungspotential zu rechnen. Zwar ist durchaus denkbar, dass Magisterabsolvent*innen der Soziologie und der Germanistik auf einem begrenzten Teil des Arbeitsmarkts konkurrieren - sollte diese Konkurrenz die Notenvergabe der Prüfenden in den Magisterprüfungen beeinflussen, sollte dies jedoch aufgrund der gemeinsamen Beeinflussung parallel geschehen.

Möglich ist hingegen, dass Unterschiede in der regionalen Arbeitslosigkeit zu Unterschieden im Notenniveau zwischen Hochschulen innerhalb eines Studiengangs beitragen. In den Wirtschaftswissenschaften findet Grözinger einen notenverschlechternden Einfluss der regionalen Arbeitslosenquote (Grözinger 2015). Für die vorliegende Stichprobe kann dies aufgrund fehlender Daten auf Hochschulebene jedoch nicht überprüft werden.

Auch der nachgewiesene negative Zusammenhang zwischen Lehrbedingungen und Noten entwickelt sich intertemporär als relationaler Effekt vorangegangener Zustände. Es ist eher unwahrscheinlich, dass die Größe der Kurse an anderen Hochschulen die Selektionsneigung von Dozierenden (Ausgleichshypothese) beeinflusst - sofern sie deren Größe überhaupt kennen und beurteilen können. Höchstens der positive Zusammenhang zwischen Prüfungszahlen als Indikator für Lehrqualität und Notenhöhe, der einen leistungskonformen Effekt markiert, könnte zur Erklärung von Querschnittsunterschieden zwischen Hochschulen im gleichen Studiengang beitragen. Da aber mit Göttingen und Berlin nur in Germanistik gerade einmal zwei Hochschulen vorliegen, für die die Daten im selben Studiengang eindeutig auf die Lehrqualitätsthese hinweisen, erübrigt sich eine entsprechende Analyse.

Gesellschaftliche Ereignisse und Entwicklungen

Im Gegensatz zur US-amerikanischen Forschung ist in der weniger umfassenden deutschen Diskussion um Notenverbesserungen bisher noch kein Verweis auf bestimmte gesellschaftliche Ereignisse oder Entwicklungen zu finden, die eine Noteninflation angestoßen oder verfestigt haben könnten. In Anbetracht der dynamischen Entwicklung des deutschen Hochschulsystems in den 1960er und 1970er Jahren soll an dieser Stelle jedoch zumindest eine Einordnung der ersten, starken Phase der Notenverbesserung (Anfang/Mitte der 1960er bis Anfang der 1970er Jahre), die sich über die Studiengänge ja relativ einheitlich darstellt, in diesen zeitlichen Kontext erfolgen:

Im Laufe der Bildungsexpansion, nach Hüther und Krücken die „auffälligste Dynamik in den westlichen Hochschulsystemen in den letzten Jahrzehnten“ (Hüther/Krücken 2016:67), verdreifachte sich die Anzahl an Dozierenden an deutschen Hochschulen zwischen 1960 und 1970, also innerhalb von 10 Jahren. Während die Anzahl wissenschaftlicher Mitarbeiter*innen danach langsamer wuchs, verdoppelten sich die Stellen für Professor*innen bis 1975 sogar noch einmal. Mit diesem Anstieg ging eine Verjüngung des Lehrkörpers und eine (zumindest kurzweilige) tendenzielle Annäherung an die Interessen und Positionen der Studierenden einher (Oehler 1986). Parallel zu dieser Erweiterung und Verjüngung des Lehrkörpers gewann die von der Studierendenbewegung flankierte wissenschaftliche Kritik an der Aussagekraft von Noten und damit an der Notengebung an sich zunehmend an Aufmerksamkeit und ließ zumindest Teile der Dozierendenschaft nicht unbeeindruckt (Ziegenspeck 1999). Dieser Erklärungsansatz für eine erste, fachübergreifende Phase der Aufweichung von Bewertungsstandards ist konform mit der aufgezeigten Notenentwicklung: Da die Magisterprüfung erst im Laufe der 1960er Jahre wieder flächendeckend eingeführt wurde (Grüner 1971) besteht für diesen Zeitraum keine Relevanz für die Notengebung in Soziologie und Germanistik Magister – zwei der drei Studiengänge, in denen (sicher) keine Verbesserung gegeben ist. Dass die Studierendenbewegung, sollte sie das Selektionsklima unter den Hochschulprüfer*innen tatsächlich beeinflusst haben, keinen Effekt auf das Notenniveau in den Rechtswissenschaften gehabt hat, könnte zum einen auf die gesetzlich verankerten Laufbahnanforderungen, zum anderen auf die große Distanz zwischen der politisch links stehenden Bewegung und dem vergleichsweise rechts einzuordnenden juristischen Lehrpersonal (vgl. Maier-Leibnitz/ Schneider 1991) zurückzuführen sein.

Einmal angestoßen, lässt sich die Notenverbesserung im Folgenden, wie gezeigt wurde, vor allem auf die unterschiedliche Elastizität hin zu Verbesserung und Verschlechterung zurückführen, die die kontinuierliche Reaktion auf die zyklische Arbeitsmarktentwicklung bzw. die gegebenen Lehrbedingungen mit sich bringt.

Zusammenfassung

Es lässt sich festhalten, dass FH1 (Mögliche Unterschiede im Notenniveau zwischen nach Fächern, Abschlüssen und/oder Hochschulen abgrenzbaren Gruppen von Prüflingen sind sowohl auf leistungskonforme, als auch auf leistungsexterne Ursachen zurückzuführen) bestätigt werden kann.

Dabei gestalten sich auch die leistungsexternen Prüfungsbedingungen in ihrem Einfluss auf die Notengebung, kann ein solcher (hier oder im Falle des Anteils Professorinnen durch Grözinger (2017)) nachgewiesen werden, weitestgehend studiengangspezifisch. Eine Ausnahme bilden die eingesetzten Prüfungsverfahren. Es zeigt sich ein genereller Effekt des Anteils mündlicher Prüfungen gegenüber der schriftlichen Hausarbeit sowie der (höheren) Gewichtung der Arbeit bei der Gesamtnotenberechnung. Er entsteht aus einer nachgewiesenen Notenhierarchie der Prüfungsverfahren: In der schriftlichen Hausarbeit werden im Mittel die besten Noten vergeben, in Klausuren die schlechtesten. Mündliche Prüfungen liegen dazwischen. Das (wieder studiengangspezifische) Ausmaß des Effekts der formalen Prüfungsbedingungen kann aufgrund der Datenstruktur nicht genau bestimmt werden[110], ein Einfluss ist jedoch gegeben.

Zweifelsfrei belegt werden kann ein Zusammenhang zwischen Prüflingsanzahl bzw. Anzahl der Studierenden und der Notenhöhe: In Studiengängen mit einheitlichem Arbeitsmarkt führen die als Folge einer Arbeitsmarktüberfüllung (eines Mangels) sinkenden (steigenden) Erstsemesterzahlen zu schlechter (besser) werdenden Noten und umgekehrt. In Biologie zeigt sich dieser Zusammenhang als Ausgleichseffekt für schlechte Lehrbedingungen. Hier reagieren die Noten direkt und negativ auf die Prüflingsanzahl, nicht auf die Anzahl Erstsemester. In Chemie und BWL reagieren die Noten ebenfalls unmittelbar auf die Lehrbedingungen, allerdings synchron zu ihnen: Verschlechtern sie sich mit einer steigenden Zahl Studierender, verschlechtern sich auch die Noten und umgekehrt. In Studiengängen mit diffusem Arbeitsmarkt lassen sich der Ausgleichseffekt und der leistungskonforme Einfluss der Lehrbedingungen auch auf Hochschulebene aufzeigen. FH4b (Signifikante, im Zeitverlauf stabile Abweichungen vom durchschnittlichen Notenniveau im jeweiligen Studiengang an einzelnen Hochschulen sind zum Teil auf hochschulspezifische Ausprägungen der Prüfungsbedingungen zurückzuführen) lässt sich spätestens durch dieses Ergebnis bestätigen. Auch FH6 (Je nach Existenz eines einheitlichen Arbeitsmarkts verlaufen Notenniveaus studiengang- oder hochschulspezifisch in Abhängigkeit von Entwicklungen des Wirtschaftssystems) trifft damit zu.

Unabhängig von der Art des Zusammenhangs zwischen Studierenden-/Prüflingsanzahl birgt dieser Zusammenhang einen Mechanismus, der die langfristige Verbesserung der Noten im Zeitverlauf be-

[110] Auch die Vermutung, dass Brüche im Notenniveau auf Änderungen der formalen Prüfungsbedingungen zurückgeführt werden können, kann aufgrund der gegebenen Datenstruktur nicht überprüft werden. FH5 (Es existieren hochschulinterne Einflüsse auf die Notengebung, die Brüche im Notenniveau einer konkreten Hochschule produzieren und es im Zeitverlauf verändern) kann damit als einzige Forschungshypothese (an dieser Stelle) nicht bestätigt werden.

wirkt: Der jeweils notenverbessernde Effekt ist stets stärker als der notenverschlechternde. Diese unterschiedliche Elastizität der Noten führt langfristig zu einer zyklisch verlaufenden Verbesserung der durchschnittlichen Abschlussnoten.
Ihren Ursprung könnte die Notenverbesserung an deutschen Hochschulen in der Bildungsexpansion der 1960er und 1970er Jahre haben: Diese Phase stellt den Hochpunkt der wissenschaftlichen Kritik an der allgemeinen Aussagekraft von Noten dar. Unter dem Eindruck der Studierendenbewegung haben der Ausbau und die Verjüngung des Lehrkörpers diese kritische Perspektive auf die Notengebung womöglich direkt in die Fakultäten getragen und damit zu einer milderen Beurteilungspraxis als zuvor geführt. Die Zusammensetzung der Lehrenden nach Alter weist keine systematischen Unterschiede in Quer- oder Längsschnitt auf (wohl aber nach Geschlecht, hier ist von einem paarvergleichsspezifischen Effekt auszugehen) und unterschiedliche Notenniveaus bilden keinen Anreiz zur Angleichung der Noten durch schlechter bewertende Prüfende.

Für die betrachteten Prüfungsbedingungen konnte in mindestens einer der vier Dimensionen Studienganglevel/Querschnitt, Studienganglevel/Längsschnitt, Hochschullevel/Querschnitt, Hochschullevel/Längsschnitt zumindest eine Einschätzung des maximalen Einflusspotentials abgegeben werden. Die wissenschaftliche Ausrichtung (rein vs. angewandt) eines Studiengangs und die damit einhergehende Ausrichtung der Nachwuchsproduktion auf den Wissenschaftssektor oder den gesamten Arbeitsmarkt dürfte im marktwirtschaftlichen System eine untergeordnete Rolle spielen - für die Auswirkung auf das Selektionsklima ist, wie gezeigt werden konnte, entscheidend ob für einen einheitlichen oder diffusen Arbeitsmarkt produziert wird.
Mögliche Einflüsse des Standardisierungsgrades von Inhalten/Prüfungen[111], der Bezugsnormenorientierung, der Häufung von Wahrnehmungsfehlern und der Finanzierungsstrukturen können hingegen aufgrund mangelnder Daten nicht genauer betrachtet werden, zur Forschungsintensität lassen sich keine aussagekräftigen Berechnungen durchführen. Dass Noten genutzt werden, um Studierende anzuziehen und dadurch finanzielle Mittel aufzustocken ist für deutsche Hochschulen aufgrund der noch geringen Kopplung von Fördermitteln an derartige Outputfaktoren jedoch (noch) nicht anzunehmen (Bauer/Grave, 2011). Auch der Konsument*innenstatus von Studierenden ist nicht ohne weiteres aus den US-amerikanischen auf deutsche Verhältnisse übertragbar, wie ein Vergleich der Bedeutung von Studiengebühren zwischen den beiden Hochschulsystemen schnell offenbart. Über die Rolle von Ideologien in den einzelnen Studiengängen ließen sich im hier genutzten Analyseschema höchstens Mutmaßungen anstellen (zumal es nach Einschätzung des Autors bereits grundsätzlich

[111] Grözinger (2017) schlussfolgert aus seinen Auswertungen der Hochschulprüfungsstatistik eine bessere Vergleichbarkeit der Noten in Fächern, in denen sich Prüfungen leichter standardisieren lassen.

fragwürdig erscheint, einzelnen Studiengängen kohärente ideologische Fundamente zuordnen zu wollen).

Dass die in der US-amerikanischen Forschung häufig thematisierten Lehrevaluationen in dieser Arbeit keine Beachtung finden, ist dadurch begründet, dass ihnen an deutschen Hochschulen im 20. Jahrhundert noch keine große Bedeutung zukam (an keiner der Hochschulen in der Stichprobe wurden vor 2004 zentral verwaltete Evaluationen zur Qualitätssicherung durchgeführt). Deshalb kann die langfristige Verbesserung der Noten seit den 1960ern hierzulande nicht mit einem unintendierten Effekt dieses Instruments der Qualitätssicherung erklärt werden.

9. Zusammenfassung und Fazit

Die Notengebung an deutschen Hochschulen ist ein bisher kaum untersuchtes Phänomen. Obwohl das Thema an Hochschulen allgegenwärtig ist, gab es bisher neben einzelnen, in ihrer Reichweite beschränkten, deskriptiven Studien nur einen einzigen Versuch, Unterschiede und Entwicklungen der Notengebung sowie Einflüsse auf sie systematisch zu erfassen (Müller-Benedict/Tsaraouha 2011). Die vorliegende Arbeit hat diesen Versuch aufgenommen und weitergeführt. Anhand erstmals verfügbarer Zeitreihendaten konnten grundlegende Muster und Dynamiken der Notengebung, verstanden als zweistufiger Prozess der Messung und Beurteilung, aufgedeckt werden

Es konnte gezeigt werden, dass fachspezifische Prüfungsbedingungen existieren, die sich zum Teil über einen langen Zeitraum zurückverfolgen lassen, weshalb sich langfristig stabile fachspezifische Muster der Notengebung erwarten lassen. Die seit 1960 vergebenen Abschlussnoten in bis zu 12 Studiengängen an sieben untersuchten Hochschulen offenbaren dabei - aufgrund des bisherigen Forschungsstands ebenfalls zu erwartende - Unterschiede in der Höhe des Notenniveaus zwischen Abschlussarten und Fächern, die eine Analyse auf Studiengang- statt auf Fachebene verlangen. Diese Analyse deckt hochschulübergreifende, zeitlich stabile Differenzen im Notenniveau auf, aus denen sich eine langfristig herrschende Notenhierarchie ergibt: Die Rangfolge beginnt mit Biologie als Studiengang mit den besten Noten, dicht gefolgt von Psychologie. In den ersten juristischen Staatsexamen werden im Durchschnitt die schlechtesten Noten vergeben, sie sind mehr als doppelt so hoch wie in Biologie. In den beiden wirtschaftswissenschaftlichen Studiengängen BWL und VWL sind die schlechtesten Durchschnittsnoten innerhalb der Diplomstudiengänge zu finden. Damit bestätigen sich die Eindrücke bisheriger Erhebungen, nach denen in den Naturwissenschaften bessere Noten als in den Geisteswissenschaften vergeben werden, während in ingenieur- und wirtschaftswissenschaftlichen Studiengängen und vor allem in den Rechtswissenschaften die schlechtesten Noten erteilt werden.

Die Notenskala wird bei der Vergabe der Abschlussnoten, wie bereits vom Wissenschaftsrat (2003) bemerkt, unterschiedlich breit bedient: Am konzentriertesten ist die Nutzung in Biologie mit 96.5% ‚sehr guten' oder ‚guten' Abschlüssen bei 0.1% ‚ausreichend' und Psychologie mit 95.0% ‚sehr guten' oder ‚guten' Abschlüssen bei 0.2% ‚ausreichend' über den gesamten Erhebungszeitraum betrachtet. Am breitesten wird das Spektrum in VWL genutzt, wo die Anteile der Extrempole 1 (4.8%) und 4 (10.0%) am ehesten ausgeglichen sind.

Die Studiengänge mit sehr guten und sehr schlechten Notenniveaus weisen eine niedrige Streuung der Noten auf, für die Studiengänge im mittleren Bereich der Notenskala zeigt sich kein Muster. Anhand der Anzahl an Jahren, in denen sich das Notenniveau zwischen den einzelnen Studiengängen signifikant unterscheidet, lassen sich aus der Rangfolge der 12 Studiengänge sieben Positionen be-

stimmen, die ein ähnliches Notenniveau aufweisen. Unter Berücksichtigung der für einzelne Zeitpunkte vorhandenen Vergleichsdaten muss in Betracht gezogen werden, dass die Stichprobendaten das Notenniveau in BWL jedoch leicht überschätzen, weshalb womöglich eher von sechs statt von sieben Positionen auszugehen ist, da das BWL-Niveau näher am VWL-Niveau liegen dürft. Die Differenzen im Notenniveau zwischen den Positionen müssen jedoch stets im zeitlichen Kontext betrachtet werden - ihre zeitliche Stabilität variiert deutlich. Die Grenze zwischen gemeinsamem und signifikant unterschiedlichem Notenniveau liegt im Bereich 0.22 bis 0.24 Noten Abstand. Insgesamt lassen sich drei unterschiedliche Beziehungsmuster zwischen den einzelnen Studiengängen ausmachen: 1. Hohe Anzahl signifikant differenter Jahre und hohe zeitliche Stabilität dieser Differenzen, 2. Niedrige Anzahl signifikant differenter Jahre und niedrige oder keine Stabilität dieser Differenzen und 3. Niedrige bis mittelhohe Anzahl signifikant differenter Jahre und niedrige bis mittlere Stabilität dieser Differenzen. Auffällig ist, dass die beiden in der Stichprobe enthaltenen Magister- und die beiden Lehramtsstudiengänge (wie auch bei Hitpass/Trosien 1987 und Wissenschaftsrat 2003; 2007; 2012) jeweils ein ähnlich hohes Notenniveau aufweisen.

Im Zeitverlauf entwickeln sich die durchschnittlichen Abschlussnoten entgegen dem Eindruck, den bisherige Publikationen zum Thema erwecken nicht nur konstant zum Besseren, sondern in zwei unterschiedlichen Verlaufsformen: Entweder sie verbessern sich tatsächlich langfristig, dann allerdings begleitet durch zyklische Schwankungen oder sie verlaufen zyklisch auf einem relativ stabilen Notenniveau. Die Diplomstudiengänge zählen mit Ausnahme von Maschinenbau (hier liegen allerdings auch nur Daten von zwei Hochschulen vor, weshalb kein aussagekräftiger Rückschluss auf das im Studiengang üblicherweise vorzufindende Notenniveau und dessen Entwicklung möglich ist) alle zur ersten Gruppe, ebenso die beiden Lehramtsstudiengänge, die neben einem ähnlichen Notenniveau im Querschnitt auch im Längsschnitt eine zeitlich parallele Entwicklung aufweisen, wenn auch mit stärkerem Trend in Deutsch. Die beiden Magisterstudiengänge und das erste juristische Staatsexamen zählen zu Letzteren. Völlig konstante Notenniveaus sind über alle Prüflinge gemittelt nicht zu finden, auch langfristige Verschlechterungen der Noten sind nicht zu beobachten.
In allen Studiengängen mit langfristiger Verbesserung setzt die erste Verbesserungsperiode in etwa zeitgleich zu Beginn/Mitte der 1960er Jahre ein und endet Anfang der 1970er Jahre, die zweite Verbesserungsphase beginnt jeweils im Laufe der 1980er Jahre, wobei sich das Ausmaß der Verbesserung und das Niveau auf dem sie sich vollzieht, studiengangspezifisch unterscheiden. Die Verbesserung der Noten wird durch Plateauphasen unterbrochen, die etwa zeitgleich einsetzen, allerdings von unterschiedlicher Dauer sind. Diese Plateauphasen treten immer dann auf, wenn der Abwärtstrend relativ schwach ist und im gleichen Zeitraum eine Aufwärtsbewegung der zyklischen Zeitreihenkomponente, die in allen Studiengängen eine Dauer von ca. 20 Jahren aufweist, stattfindet. In BWL und

VWL, den Studiengängen mit dem schlechtesten Notenniveau der Studiengänge mit langfristiger Verbesserung nimmt die Stärke des Abwärtstrends im Zeitverlauf zu. In Psychologie und Biologie ist die Verbesserung nach kurzer, starker Dynamik bereits nach der ersten Phase weitestgehend abgeschlossen. Seit Beginn der 1970er Jahre kann dort von grade compression gesprochen werden. Unabhängig vom studiengangspezifischen Notenniveau zeigt sich ein Zusammenhang zwischen dem Verlauf der Durchschnittsnoten und der zugehörigen Streuung, welche sich mit dem Absinken des Notenniveaus verringert, wodurch in den betroffenen Studiengängen die Differenzierung von Leistung zunehmend schwieriger geworden ist.
In den Studiengängen mit zyklischem Verlauf auf relativ stabilem Notenniveau verlaufen die Noten in unterschiedlichen Schwankungsbreiten, über die Jahre hinweg streuen sie am stärksten in Soziologie, am schwächsten in Jura. Die Auf- und Abwärtsbewegungen der Noten beginnen und enden leicht versetzt, die Zyklen dauern aber wie auch in den Studiengängen mit Verbesserung ca. 20 Jahre.

Erklärungsbedürftige Differenzen im Notenniveau und mehr oder weniger stabile Unterschiede in der langfristigen Entwicklung der Noten existieren jedoch nicht nur zwischen den Studiengängen. Auch innerhalb der Studiengänge zeigen sich aus der parallel zur fachkulturellen Sozialisation erfolgenden Einbindung der Prüfenden in ein hochschulspezifisches Setting aus lokalen Prüfungssystemen und lokaler Prüfer*innengemeinschaft theoretisch ableitbare hochschulspezifische Abweichungen von Studiengangniveau und -entwicklung. Wie auch eine genauere Betrachtung der vorhandenen Vergleichsstudien bereits erwarten ließ, verlaufen innerhalb desselben Studiengangs die langfristigen Verbesserungen nicht unbedingt auf dem gleichen Niveau und in gleichem Ausmaß, setzen die Verbesserungen nicht überall gleichzeitig ein und verlaufen die Zyklen teils in unterschiedlich langer Dauer und in unterschiedlichen Notenhöhen. Die Streuung der Noten fällt im selben Studiengang ebenfalls teils unterschiedlich stark aus. Gemeinsamkeiten finden sich in den Hochschulen vor allem hinsichtlich der langfristigen Entwicklung: In der Regel sind auch an den einzelnen Hochschulen zyklische Bewegungen von 10-20 Jahren Länge zu finden, die entweder einen langfristigen Abwärtstrend begleiten oder sich relativ gleichmäßig um ein konstantes Notenniveau herum bewegen. Die Abwärtstrends weisen in den meisten Studiengängen mit langfristiger Verbesserung einen einigermaßen parallelen Zeitraum auf, in dem der Trend am stärksten ist. In den meisten Studiengängen überwiegt zudem die Anzahl der Hochschulen, die den gleichen Notenverlauf aufweisen wie der Gesamtdurchschnitt. Lediglich in Biologie und Chemie ist das Verhältnis mit je vier Standorten mit langfristigem Abwärtstrend und drei ohne einigermaßen ausgeglichen.

Zur Erklärung der festgestellten Unterschiede zwischen Abschlüssen, Fächern und Hochschulen sowie der zeitlichen Entwicklung der Noten können, wie die Systematisierung der erarbeiteten potentiellen

Einflussfaktoren zeigt, grundsätzlich zwei Hauptdimensionen herangezogen werden: Leistungskonforme und leistungsunabhängige Faktoren. Es lassen sich innerhalb der leistungsexternen Dimension dabei zwei Typen von Prüfungsbedingungen ableiten: Einerseits einmalig wirkende Eingriffe in das Prüfungsgeschehen, andererseits kontinuierlich wirksame Langzeiteinflüsse. Leistungskonforme Prüfungsbedingungen sind hingegen stets kontinuierlich präsent.

Von Letzteren stellen die Eingangseignung der Studierenden (wie der Stand der Forschung erwarten ließ) sowie die Lehrbedingungen relevante Größen in ihrem Einfluss auf die Notengebung dar. Möglicherweise besteht auch ein begrenzter Einfluss der sozialen Herkunft sowie der Studiendauer, die als kompositionelle Faktoren durchschnittliche Notenniveaus beeinflussen könnten - genau lässt sich dies anhand der vorhandenen Informationen nicht sagen. Allerdings wären entsprechende Effekt als vergleichsweise gering einzuschätzen und auf bestimmte Paarvergleiche beschränkt, während die erklärungsbedürftigen Differenzen zwischen anderen Paarvergleichen noch erhöht würden, was eine Einstufung als systematische Einflussfaktoren hinfällig werden lässt (Lediglich im Zeitverlauf *könnte* die soziale Herkunft einen dauerhaft wirksamen, in seiner Stärke allerdings ebenfalls deutlich limitieren Einfluss besitzen). Die Geschlechtskomposition hat nachweislich einen nicht unbedeutenden Effekt auf die Notenhöhe, welcher jedoch ebenfalls nicht als systematisch begriffen werden kann. Die übrigen kompositionellen Studierendenmerkmale besitzen kein Erklärungspotential in Quer- und/oder Längsschnitt. Auch für die Selbstselektionsthese lassen sich keinerlei überzeugende Hinweise finden. In Bezug auf Teilprüfungen wäre hier allerdings weitere Forschung zu Modulprüfungen im BA/MA System interessant, da die Wahlmöglichkeiten dort in der Regel deutlich größer sind, die Noten noch transparenter vergeben werden.

Innerhalb der leistungsexternen Gruppe von Prüfungsbedingungen zeigt sich, dass die Wahl der Prüfungsverfahren sowie deren Gewichtung einen nennenswerten Einfluss auf die Notengebung besitzen. Auch Änderungen der formalen Prüfungsbedingungen weisen wie in US-amerikanischen Studien Effekte auf die Notenhöhe auf. Auch wenn deren Stärke sich anhand der vorliegenden Informationen nicht genau bestimmen lässt, zeigen diese beiden Befunde, dass die Notengebung durch lokale Rahmenbedingungen gerahmt wird und kein willkürliches Produkt unstrukturierter Bewertungsprozesse darstellt, wie es die Einstufung von Hochschulen als „organisierte Anarchien“ (Cohen et al. 1972:1) befürchten lässt. Wie willkürlich die hochschulspezifische Ausgestaltung dieser Rahmenbedingungen innerhalb eines Studiengangs ausfällt, ist dagegen eine andere Frage.

Neben diesen einmalig wirkenden Faktoren, die den langfristigen Trend zur Notenverbesserung nicht erklären können, kann zudem ein langfristiger Zusammenhang zwischen der Entwicklung der Studierenden- bzw. Prüflingsanzahl und der Notenentwicklung nachgewiesen werden, auf den Müller-Benedict und Tsarouha (2011) bereits Hinweise fanden: In Studiengängen mit einheitlichem Arbeitsmarkt führen die als Folge einer Arbeitsmarktüberfüllung (eines Mangels) sinkenden (steigenden)

Erstsemesterzahlen zu schlechter (besser) werdenden Noten und umgekehrt. In Biologie zeigt sich dieser Zusammenhang als Ausgleichseffekt für schlechte Lehrbedingungen, in BWL und Chemie verbessern bzw. verschlechtern sich die Noten parallel zu den Lehrbedingungen.
Die Zusammensetzung der Lehrenden nach Alter weist keine systematischen Unterschiede in Quer- oder Längsschnitt auf (wohl aber nach Geschlecht, hier ist von einem paarvergleichsspezifischen Effekt auszugehen) und unterschiedliche Notenniveaus bilden keinen Anreiz zur Angleichung der Noten durch schlechter bewertende Prüfende. Aufgrund mangelnder Daten bzw. Berechnungsproblemen konnten mögliche Einflüsse des Standardisierungsgrades von Inhalten/Prüfungen, der Bezugsnormenorientierung, der Häufung von Wahrnehmungsfehlern, der Forschungsintensität und der Finanzierungsstrukturen nicht analysiert werden. Für letztere ist allerdings wie für die wissenschaftliche Ausrichtung eines Fachs (rein vs. angewandt) im deutschen Hochschulsystem nicht von einem relevanten Effekt auf die Notengebung auszugehen.

Der in der Mehrzahl der untersuchten Studiengänge festzustellende Trend zu besseren Noten entsteht durch einen Mechanismus ungleicher Elastizität in der Notenvergabe: In Phasen der Verbesserung werden die Noten deutlich besser als sie sich in Phasen der Verschlechterung verschlechtern. Ein möglicher Erklärungsansatz für die Entstehung von Notenverbesserungen findet sich in der Bildungsexpansion der 1960er und 1970er Jahre, während der die wissenschaftliche Kritik an der Notengebung über die Studierendenbewegung sowie den Ausbau inklusive Verjüngung des Lehrkörpers Einzug in die Fakultäten erhalten und die Beurteilungspraxis abgemildert haben könnte.

Erwartungsgemäß stellt sich die Notengebung an deutschen Hochschulen also als komplexes Phänomen dar. Fachkulturell bedingte Unterschiede in Niveau und Entwicklung der Noten überschneiden sich mit hochschulspezifischen Ausgestaltungen der studiengangüblichen Muster. Um die aufgezeigten Differenzen im Notenniveau und Verbesserungsprozesse auf den unterschiedlichen Aggregatebenen erklären zu können, greift der reine Verweis auf Leistungsunterschiede in Quer- bzw. Längsschnitt ebenso wie unterschiedliche bzw. im Zeitverlauf gesunkene Bewertungsstandards zu kurz: Sowohl leistungskonforme als auch leistungsexterne Prüfungsbedingungen bestimmen die Notengebung mit.
Dabei zeigt sich einerseits, dass weniger Faktoren relevante Auswirkungen auf die Notengebung besitzen, als es die öffentliche Debatte und auch einige wissenschaftliche Publikationen nahelegen - vor allem Indizien für die intendierte Nutzung von Noten als Steuerungsinstrument (die Hochschule bzw. der Fachbereich/das Institut als Akteur) lassen sich nicht finden. Andererseits wird deutlich, dass die uneingeschränkte Vergleichbarkeit von Abschlussnoten in ihrer bisherigen absoluten Interpretationsweise trotz einer relativ geringen Anzahl an Einflussfaktoren eine Illusion bleiben muss. Denn

erstens muss akzeptiert werden, dass neben systematischen Effekten auch ein erheblicher Teil unsystematischer leistungskonformer Einflüsse auf die Notenvergabe besteht, der sich einer gezielten Steuerung entzieht (wobei sowieso fraglich wäre, wie sinnvoll ein Eingriff in leistungskonforme Einflüsse wäre, welche ja keine Leistungsverzerrungen bewirken). Und zweitens muss akzeptiert werden, dass auch die nachgewiesenen systematischen Einflussfaktoren sich zum Teil nicht sinnvoll regulieren, geschweige denn standardisieren lassen (z.B. die Auswirkungen der zyklischen Entwicklung der Studierendenzahlen).

Versuche der vollständigen Vereinheitlichung von Notenniveaus wären deshalb bereits von vorneherein zum Scheitern verurteilt. Auch simple Korrekturversuche unerwünschter Dynamiken, wie die künstliche Anhebung von Notenniveaus bei drohender Notenkompression am unteren Ende der Notenskala, die im Rahmen der Bachelor-Einführung vollzogen wurde (Grözinger (2017) bzw. McGrory (2017) für das Lehramt), werden die Vergleichbarkeit von Noten (langfristig) nicht wieder erhöhen, solange die Ursachen für diese Dynamiken bestehen bleiben und die Streuung der Noten nicht wieder steigt - vor allem nicht, wenn die Noten im Master wieder sinken und damit die Charakteristik der alten Studienabschlüsse repliziert wird, der Bachelor nun quasi den höheren Selektionsgrad der Zwischenprüfung (wenn auch immerhin mit akademischen Abschluss), der Master den niedrigeren Grad der alten Abschlussprüfung erhält.

Während sich die Notendifferenzen in Quer-und Längsschnitt auf Studiengangebene in Karrieren mit homogenem Arbeitsmarkt als unvermeidbar aber auch nicht weiter folgenreich auffassen lassen, sind die Auswirkungen bei einem geteilten Arbeitsmarkt und vor allem innerhalb eines Studiengangs enorm: Jede Verwendung der Abschlussnote als Leistungsindikator ist zwangsläufig an zahlreiche diskrete Vergleichseinheiten gebunden, deren vollständige Matrix dabei bestenfalls unter Verwendung der Hochschulprüfungsstatistik abrufbar ist. Und die institutionellen, hochschulspezifischen Prüfungsbedingungen, die diese Differenzen auf Hochschulebene produzieren, werden sich in Zukunft aufgrund der aktuellen hochschulpolitischen Entwicklungen (Stichwort: Profilbildung der Hochschulen) eher weiter ausdifferenzieren als vereinheitlichen - auch wenn eine Vereinheitlichung zumindest in Bezug auf die formalen Prüfungsbedingungen und Gewichtungskriterien von Prüfungsleistungen wünschenswert wäre und den Einfluss leistungsexterner Verzerrungen bereits reduzieren würde. Da nicht verhindert werden kann, dass Absolvent*innen des gleichen Studiengangs ungleichen, standortspezifischen Bewertungsstandards unterliegen und auch nicht alle studiengangspezifischen Notenniveaus öffentlich bekannt sind, sollte an den deutschen Hochschulen akzeptiert werden, dass die Notengebung selbst durch epistemologische, wissenschaftsorganisatorische und kulturelle Faktoren derart verzerrungsbelastet ist, dass in erster Linie nicht die Notenvergabe, sondern die Noteninterpretation einer Reform bedarf. Hier sind relationale Modelle denkbar, die Notendurchschnitte in mehrere Kontexte einordnen (wie der Wissenschaftsrat (2012) es bereits ansatzweise

empfiehlt, es an einigen Hochschulen auch bereits ansatzweise praktiziert und zuletzt auch von der Kultusministerkonferenz im Bologna-Reformpapier empfohlen wird): Noten könnten im Querschnitt etwa in Relation zu den Notendurchschnitten im gesamten Studiengang an der jeweiligen Hochschule oder auch auf Bundesebene, im Längsschnitt zu den Notendurchschnitten vergangener Semester (einzeln oder im Aggregat) gesetzt werden. Eine flächendecke Etablierung eines derartigen, einheitlich gehandhabten Systems könnte die heute stark eingeschränkte Aussagekraft von Examensnoten wieder erhöhen.

Literaturverzeichnis

Achen, Alexandra C./ Courant, Paul N. (2009): What Are Grades Made Off? In: The Journal of Economic Perspectives. Vol. 23 (3), S. 77-92.

Adelman, Clifford (1995): The New College Course Map and Transcript Files: Changes in Course-Taking and Achievement, 1972-1993. Washington, DC : Office of Educational Research and Improvement, US Department of Education.

Adelman, Clifford (2008): Undergraduate Grades: A More Complex Story Than "Inflation". In: Hunt, L.H. (Hrsg.): Grade Inflation. Academic Standards in Higher Education. New York : State University of New York Press. S. 13-44.

Agnew, Eleanor (1995): Rigorous Grading Does Not Raise Standards: It Only Lowers Grades. In: Assessing Writing. Vol. 2 (1), S. 91-103.

Anglin, Paul M./ Meng, Ronald (2000): Evidence on Grades and Grade Inflation at Ontario's Universities. In: Canadian Public Policy - Analyse De Politiques. Vol. 26 (3), S. 361-368.

Apel, Helmut (1989): Fachkulturen und studentischer Habitus : eine empirische Vergleichsstudie bei Pädagogik- und Jurastudierenden. In: Zeitschrift für Sozialisationsforschung und Erziehungssoziologie. Vol. 9 (1), S. 2-22.

Apenburg, Eckhard/Günther, Kurt/Reither, Franco (1976): Prüfungsergebnisse für den Zeitraum WS 1972/1973 bis SS 1975 für acht ausgewählte Fächer an der Universität des Saarlandes. Schriftenreihe Saarbrücker Studien zur Hochschulentwicklung. Saarbrücken : Universität des Saarlandes.

Apenburg, Eckhard/Jurecka, Peter/Tausendfreund, Regina (1977a): Studium und Lehre aus der Sicht von Lehrenden und Lernenden. Bericht über Befragungen an Studierenden und Hochschullehrern in acht ausgewählten Fächern an der Universität des Saarlandes. Saarbrücken : Universität des Saarlandes.

Apenburg, Eckhard/Grosskopf, Ruth/Schlattmann, Hartmut (1977b): Orientierungsprobleme und Erfolgsbeeinträchtigung bei Studierenden. Teil B: Ergebnisse in Tabellenform. Saarbrücken : Universität des Saarlandes.

Armingeon, Klaus (2001): Fachkulturen, soziale Lage und politische Einstellungen der Studierenden der Universität Bern. Bern: Institut für Politikwissenschaft.

Arnold, Markus (2004): Disziplin und Initiation. Die kulturellen Praktiken der Wissenschaft. In: Arnold, M./Fischer, R. (Hrsg.): Disziplinierungen. Kulturen der Wissenschaft im Vergleich. Wien : Verlag Turia + Kant. S. 18-52.

Austin, Ann E. (1990): Faculty Cultures, Faculty Values. In: New Directions for Institutional Research. Vol. 68, S.61-74.

Babcock, Phillip (2010): Real Costs of Nominal Grade Inflation? New Evidence from Student Course Evaluations. In: Economic Inquiry. Vol. 48 (4), S. 983-996.

Bagues, Manuel/Labini, Mauro S./Zinovyeva, Natalia (2008): Differential Grading Standards and University Funding: Evidence from Italy. In: CESifo Economic Studies. Vol. 54 (2), S. 149-176.

Baird, Matthew (2009): Dynamic Estimation of the Incentive Schemes and Signalling Costs of Grade Inflation. On-Line Working Paper CCPR-2009-015, California Center for Population Research, UC Los Angeles.

Bar, Talia/Kadiyali, Vrinda/Zussman, Asaf (2009): Grade Information and Grade Inflation: The Cornell Experiment. In: The Journal of Economic Perspectives. Vol. 23 (3), S. 93-108.

Bar, Talia/Zussman, Asaf (2012): Partisan Grading. In: American Economic Journal: Applied Economics. Vol. 4 (1), S.30-48.

Bargel, Tino (1988): Wieviele Kulturen hat die Universität? Ein Vergleich der Rollen- und Arbeitskultur in vierzig Einzelfächern. Hefte zur Bildungs- und Hochschulforschung (2). Konstanz : Arbeitsgruppe Hochschulforschung.

Barth, Michael M./Liu, Jun/Wells, William H. (2009): A Comparative Analysis of Grading Practices by Discipline within a College of Business. In: Academy of Educational Leadership Journal. Vol. 13 (4), S.93-107.

Barz, Andreas/Miethig, Thomas (1993): Qualität der Lehre. Ein Politikum auf dem Prüfstand - Ergebnisse der Professorenumfrage an der Universität Kaiserslautern. Kaiserslautern.

Bauer, Thomas K./Grave, Barbara S. (2011): Performance-related Funding of Universities: Does More Competition Lead to Grade Inflation? IZA Discussion Paper No.6073.

Bayer, Klaus (2013): Immer bessere Noten? Über die Zerstörung der geisteswissenschaftlichen Prüfungskultur. In: Forschung & Lehre. Vol. 1/13. S. 36-38.

Bearden, James/Wolfe, Raymond N. (2001): Reaction to Compton and Metheny (2000): “Assessment of Grade Inflation in Higher Education”. In: Perceptual and Motor Skills. Vol. 92, S.263-264.

Becher, Tony (1981): Towards a definition of disciplinary cultures. In: Studies in Higher Education. Vol. 6 (2), S. 109-122.

Becher, Tony (1984): The Cultural View. In: Clark, B.R. (Hrsg.): Perspectives on Higher Education. Eight Disciplinary and Comparative Views. Berkeley : University of California Press. S. 165-198.

Becher, Tony (1987a): Disciplinary discourse. In: Studies in Higher Education. Vol. 12 (3), S. 261-274.

Becher, Tony (1987b): The Disciplinary Shaping of the Profession. In: Clark, B.R. (Hrsg.): The Academic Profession. National, Disciplinary and Institutional Settings. Berkeley : University of California Press. S. 271-303.

Becher, Tony (1989): Academic Tribes and Territories. Intellectual Enquiry and the Cultures of Disciplines. Milton Keynes : Open University Press.

Becher, Tony (1990): The Counter-culture of Specialisation. In: European Journal of Education. Vol. 25 (3), S. 333-346.

Becker, Birgit (2010): Bildungsaspirationen von Migranten. Determinanten und Umsetzung in Bildungsergebnisse. Arbeitspapiere - Mannheimer Zentrum für Europäische Sozialforschung, Vol. 137.

Becker, Rolf (2011): Integration von Migranten durch Bildung und Ausbildung - theoretische Erklärungen und empirische Befunde. In: Becker, R. (Hrsg.): Integration durch Bildung - Bildungserwerb von jungen Migranten in Deutschland. Wiesbaden: VS Verlag für Sozialwissenschaften. S. 11-38.

Behr, Andreas/Theune, Katja (2016): The causal effect of off-campus work on time to degree. In: Education Economics. Vol. 24 (2), S.189-209.

Bejar, Isaac I./Blew, Edwin O. (1981): Grade Inflation and the Validity of the Scholastic Aptitude Test. In: American Educational Research Journal. Vol. 18 (2), S. 143–156.

Bernstein, Basil (1977): Beiträge zu einer Theorie des pädagogischen Prozesses. Frankfurt a.M. : Suhrkamp.

Betz, Dieter (1974): Rhythmische Schwankungen als Fehler in der Notengebung bei mündlichen Prüfungen. In: Psychologie in Erziehung und Unterricht. Vol. 21, S. 1-14.

Biglan, Anthony (1973): The Characteristics of Subject Matter in different Academic Areas. In: Journal of Applied Psychology. Vol. 57 (3). S. 195-203.

Birkel, Peter (1978): Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung. Bochum : Kamp.

Birnbaum, Robert (1977): Factors Related to University Grade Inflation. In: The Journal of Higher Education. Vol. 48 (5), S. 519-539.

Bitz, Ferdinand (1989): Empirisch-vergleichende Studie zur Entwicklung der Diplomprüfung in Physik und Psychologie an den Hochschulen der Bundesrepublik Deutschland. Bonn : Rheinische Friedrich-Wilhelms-Universität.

Böhmer, Susan/Neufeld, Jörg/Hinze, Sybille/Klode, Christian/Hornbostel, Stefan (2011): Wissenschaftler-Befragung 2010: Forschungsbedingungen von Professorinnen und Professoren an deutschen Universitäten. IFQ-Working Paper 8.

Borchert, Karlheinz (1986): Soziale Herkunft und berufliche Karriere. Ergebnisse einer Absolventenbefragung an Fachhochschulen. In: Zeitschrift für Sozialisationsforschung und Erziehungssoziologie. Vol. 6 (1), S. 111-128.

Boretz, Elizabeth (2004): Grade inflation and the myth of student consumerism. In: College Teaching. Vol. 52 (2), S. 42-46.

Bourdieu, Pierre (1983): Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In: Kreckel, R. (Hrsg.): Soziale Ungleichheiten. Göttingen : Schwartz. S. 183-198.

Bourdieu, Pierre (1985): Sozialer Raum und „Klassen“. Zwei Vorlesungen. Frankfurt a.M. : Suhrkamp.

Bourdieu, Pierre (1988): Homo Academicus. Stanford : Stanford University Press.

Bourdieu, Pierre (1998): Practical Reason. On the Theory of Action. Stanford : Stanford University Press.

Bourner, Jill/Bourner, Tom (1985): Degrees of Success in Accounting. In: Studies in Higher Education. Vol. 10 (1), S. 55-68.

Brandt, Patrick T./Williams, John T. (2007): Multiple Time Series Models. Thousand Oaks, CA : Sage.

Breland, Hunter M. (1976): Grade Inflation and Declining SAT Scores: A Research Viewpoint. Paper presented at the Annual Meeting of the American Psychological Association in Washington D.C..

Brighouse, Harry (2008): Grade Inflation and Grade Variation: What's All the Fuss About? In: Hunt, L.H. (Hrsg.): Grade Inflation. Academic Standards in Higher Education. New York : State University of New York Press. S. 73-91.

Brinkmann, Gerhard (1967): Die Prognose des Studienerfolgs. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie. Vol. 19, S. 322-333.

Brunsson, Nils/Sahlin-Andersson, Kerstin. (2000): Constructing Organizations. The Example of Public Sector Reform. In: Organization Studies. Vol. 21 (4), S. 721-746.

Bundesministerium für Bildung und Forschung (2015): Das Deutschlandstipendium: Entwicklung und Verteilung.
http://www.deutschlandstipendium.de/_media/150701_Entwicklung_Verteilung_A4_4C.pdf

Bülow-Schramm, Margret (2014): Durchlässigkeit als Zielmarke für Übergänge im Hochschulsystem? Zur Ambivalenz von Heterogenität und Homogenität in der Hochschule. In: Banscherus, Ulf (Hrsg.): Übergänge im Spannungsfeld von Expansion und Exklusion : eine Analyse der Schnittstellen im deutschen Hochschulsystem. Bielefeld : Bertelsmann.

Centra, John A. (1993): Reflective Faculty Evaluation. San Francisco: Jossey-Bass.

Chan, William/Hao, Li/Suen, Wing (2007): A Signaling Theory of Grade Inflation. In: International Economic Review. Vol. 48 (3), S. 1065-1090.

Chapman, Keith (1994): Variability of Degree Results in Geography in United Kingdom Universities 1973-90: preliminary results and policy implications. In: Studies in Higher Education. Vol.19 (1), S. 89-102.

Chapman, Keith (1997): Degrees of difference: variability of degree results in UK universities. In: Higher Education. Vol. 33, S. 137–153.

Cheong, Kwang Soo (2000): Grade Inflation at the University of Hawaii-Manoa. Working Paper No. 00-2.

Clark, Burton R. (1963): Faculty Culture. In: Lunsford, T.F. (Hrsg.): The study of campus cultures. Boulder, CO: Western Interstate Commission for Higher Education. S. 39–54.

Clark, Burton R./Trow, Martin (1966): The Organizational Context. In: Newcomb, T.M./Wilson, E.K. (Hrsg.): College Peer Groups. Problems and Prospects for Research. Chicago : Aldine Publishing Company. S. 17-70.

Clark, Burton R. (1985): Listening to the Professoriate. In: Change. Vol. 17 (5), S. 36-43.

Clark, Burton R. (1987): The Academic Life. Small Worlds, Different Worlds. Princeton : The Carnegie Foundation for the Advancement of Teaching.

Cleveland, William S. (1979): Robust Locally Weighted Regression and Smoothing Scatterplots In: Journal of the American Statistical Association. Vol. 74 (368), S. 829-836.

Cochrane, Donald/ Orcutt, Guy H. (1949): Application of Least Squares Regression to relationships containing auto-correlated error terms. In: Journal of the American Statistical Association. Vol. 44 (245), S. 32-61.

Cluskey, G. R. Jr./Ehlen, Craig R./Griffin, Nathan (1997). Accounting grade inflation. In: Journal of Education for Business. Vol. 72 (5), S. 273-277.

Cohen, Michael D./March, James G./Olsen, Johan. P. (1972): A Garbage Can Model of Organizational Choice. In: Administrative Science Quarterly. Vol. 17, S. 1-25.

Compton, David M./Metheny, Brenda (2000): An Assessment of Grade Inflation in Higher Education. In: Perceptual and Motor Skills. Vol. 90, S. 527-536.

Connolly, Kevin J./Smith, Peter K. (1986): What makes a "good" degree: Variations between different departments. In: Bulletin of the British Psychological Society. Vol. 39, S. 48-51.

Correa, Hector (2001): A game theoretic analysis of faculty competition and academic standards. In: Higher Education Policy. Vol. 14, S. 175-182.

Crowl, Thomas K. (1984): Grading Behavior and Teachers' Need for Social Approval. In: Education. Vol. 104 (3), S.291-295.

Davis, James A. (1966): The campus as a frog pond: An application of the theory of relative depravation to career decisions of college men. In: American Journal of Sociology. Vol. 72, S. 17–31.

De Paola, Maria (2008): Are easy grading practices induced by low demand? Evidence from Italy. MPRA Paper No. 14425.

Dickson, Vaughan A. (1984): An Economic Model of Faculty Grading Practices. In: Journal of Economic Education. Vol. 15 (3), S. 197-203.

Diefenbach, Heike (2010): Kinder und Jugendliche aus Migrantenfamilien im deutschen Bildungssystem - Erklärungen und empirische Befunde. Wiesbaden : VS Verlag für Sozialwissenschaften.

Dippelhofer-Stiem, Barbara (1983): Hochschule als Umwelt: Probleme der Konzeptualisierung, Komponenten des methodischen Zugangs und ausgewählte empirische Befunde. Weinheim : Beltz.

Doss, D. Adrian/Pitts, Sarah T./Kamery, Rob H. (2005a): The Effects of Part-Time vs. Full-Time Teaching on Final Grades in the Introduction to Microcomputers and Business Course. In: Proceedings of the Academy of Educational Leadership. Vol. 10 (2), S. 13-18.

Doss, D. Adrian/Pitts, Sarah T./Kamery, Rob H. (2005b): The Effect of Part-Time Instruction on Final Grades in the Fundamentals of Algebra Course? In: Proceedings of the Academy of Educational Leadership. Vol. 10 (2), S. 19-24.

Doss, D. Adrian/Pitts, Sarah T./Kamery, Rob H. (2005c): What are the Effects of Part-Time Instruction on Final Grades in the Principles of Management and Organizational Behavior Course? In: Proceedings of the Academy of Educational Leadership. Vol. 10 (2), S. 25-30.

Doss, D. Adrian/Pitts, Sarah T./Kamery, Rob H. (2005d): What are the Effects of Part-Time Teaching on Grades in the Developmental English Composition Course at Private University? In: Proceedings of the Academy of Educational Leadership. Vol. 10 (2), S. 37-42.

Doss, D. Adrian/Pitts, Sarah T./Kamery, Rob H. (2005e): What are the Effects of Part-Time Instruction on Grades in the Legal Environment of Business course? In: Proceedings of the Academy of Educational Leadership. Vol. 10 (2), S. 75-80.

Doss, D. Adrian/Pitts, Sarah T./Kamery, Rob H. (2005f): What are the Effects of Part-Time Instruction on Final Grades in the Business Law Course? In: Proceedings of the Academy of Educational Leadership. Vol. 10 (2), S. 81-86.

Eaton, B. Curtis/Eswaran, Mukesh (2008): Differential Grading Standards and Student Incentives. In: Canadian Public Policy. Vol.34 (2), S. 215-236.

Eiszler, Charles F. (2002): College Students‘ Evaluations of Teaching and Grade Inflation. In: Research in Higher Education. Vol. 43 (4), S. 483-501.

Elias, K.S. (2003): Tough job market forces up grades. The Times Higher Education Supplement vom 5. Dezember 2003.

Enders, Jürgen/Teichler, Ulrich (1995): Berufsbild der Lehrenden und Forschenden an Hochschulen. Ergebnisse einer Befragung des wissenschaftlichen Personals an westdeutschen Hochschulen. Bonn : Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

Engler, Steffani (1993): Fachkultur, Geschlecht und soziale Reproduktion. Eine Untersuchung über Studentinnen und Studenten der Erziehungswissenschaft, Rechtswissenschaft, Elektrotechnik und des Maschinenbaus. Blickpunkt Hochschuldidaktik, Band 92. Weinheim : Deutscher Studien Verlag.

Erdel, Barbara (2010): Welche Determinanten beeinflussen den Studienerfolg? Nürnberg : Berichte / Universität Erlangen-Nürnberg.

Esser, Hartmut (2001): Integration und ethnische Schichtung. Arbeitspapiere - Mannheimer Zentrum für Europäische Sozialforschung. Vol. 40.

Faber, Günter/Billmann-Mahecha (2010): Praxis der Notengebung. Probleme, Erfordernisse und Möglichkeiten aus pädagogisch-psychologischer Sicht. In: Lernchancen. Vol. 74, S. 30-33.

Falkenberg, Steve (1996): Grade Inflation. http://people.eku.edu/falkenbergs/grdinfla.htm

Field, Andy (2013): Discovering statistics using IBM SPSS statistics : and sex and drugs and rock 'n' roll. Los Angeles : Sage.

Foster, David/Foster, Edith (1998): It’s a buyer’s market: “Disposable professors,” grade inflation, and other problems. In: Academe, Vol. 84 (1), S. 28–35.

Frank, Andrea (1990): Hochschulsozialisation und akademischer Habitus. Eine Untersuchung am Beispiel der Disziplinen Biologie und Psychologie. Blickpunkt Hochschuldidaktik, Band 87. Weinheim : Deutscher Studien Verlag.

Franz, Wan-Ju I. (2010): Grade inflation under the threat of students’ nuisance: Theory and evidence. In: Economics of Education Review. Vol. 29, S. 411-422.

Freeman, Donald G. (1999): Grade Divergence as a Market Outcome. In: Journal of Economic Education. Vol. 30 (4), S. 344-351.

Frey, B. S. (2008): Evaluitis - eine neue Krankheit. In: Simon, D./Matthies, H. (Hrsg.): Wissenschaft unter Beobachtung. Effekte und Defekte von Evaluationen. Wiesbaden. S. 125-140.

Friedmann, Jan/ Hinrichs, Per (2007): Inflation der Kuschelnoten. In: Unispiegel vom 14.05.2007.

Friedmann, Jan (2012): Wissenschaftsrat beklagt zu gute Noten an Unis. In: Unispiegel vom 09.11.2012.

Gaens, Thomas (2013): Von einem, der auszog, einen Leistungsindikator zu erheben. Durchfallquoten und die Problematik ihrer Bildung. In: Das Hochschulwesen. Vol. 6/2013, S. 200-206.

Gaens, Thomas (2015): Noteninflation an deutschen Hochschulen - Werden die Examensnoten überall immer besser? In: Beiträge zur Hochschulforschung. Vol. 4/2015, S.8-35.

Gaff, Jerry, G./Wilson, Robert, C. (1971): Faculty Cultures and Interdisciplinary Studies. In: The Journal of Higher Education. Vol.42 (3), S. 186-201.

Gaff, Jerry, G./Crombag, Hans, F.M./Chang, Ten, M. (1977): Environments for Learning in a Dutch University. In: Higher Education. Vol.5 (3), S. 285-299.

Gamson, Zelda (1967): Performance and Personalism in Student-Faculty Relations. In: Sociology of Education. Vol. 40 (4), S. 279-301.

Gass, Thomas/Meister, Gerhard (1996): Soziale Herkunft, Fachhabitus und Berufsantizipation : zur Soziologie der Fächer BWL und Germanistik. Ein Vergleich zweier Fachkulturen an der Universität Bern anhand von acht Interviews. Bern : Institut für Soziologie.

Gawatz, Reinhard (1991): Studium, Wissenschaft, Beruf: Berufliche Studienperspektiven westdeutscher Studierender und ihr Stellenwert für die Studienbewältigung und Studiensituation. Konstanz : Hartung-Gorre.

Geisinger, Kurt F. (1979): A Note on Grading Policies and Grade Inflation. In: Improving College and University Teaching. Vol. 27 (3), S. 113-115.

Georg, Werner (2005): Studienfachwahl: soziale Reproduktion oder fachkulturelle Entscheidung. In: ZA-Information 57. S. 61-82.

Gerholm, Tomas (1990): On Tacit Knowledge in Academia. In: European Journal of Education. Vol. 25 (3), S. 263-271.

Giese, Stefanie/Otte, Franziska/Stoetzer, Matthias-Wolfgang/Berger, Christian (2013): Erfolgreich studieren in betriebswirtschaftlichen Studiengängen. Eine empirische Analyse der Einflussfaktoren. In: Die Hochschule. Vol. 2/2013, S.40-55.

Gleich, Johann M./Meran, Georg/Bargel, Tino (1982): Studenten und Hochschullehrer: Eine empirische Untersuchung an baden-württembergischen Universitäten. Villingen-Schwenningen : Neckar-Verlag.

Gohmann, Stephan F./McCrickard, Myra J. (2001): Tenure Status and Grade Inflation: A Time Series Approach. In: Journal of the Academy of Business Education. Vol. 2 (2), S. 1-8.

Goldman, Roy D./Hewitt, Barbara N. (1975): Adaption-Level as an Explanation for Differential Grading Standards in College Grading. In: Journal of Educational Measurement. Vol. 12 (3), S. 149-161.

Goldman, Roy D./Widawski, Mel H. (1976): A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. In: Educational and Psychological Measurement. Vol. 36 (2), S. 381-390.

Gouldner, Alwin, W. (1957): Cosmopolitans and Locals: Toward an Analysis of Latent Social Roles. In: Administrative Science Quarterly. Vol. 2 (3), S. 281-306.

Grotheer, Michael et al. (2012): Hochqualifiziert und gefragt. Ergebnisse der zweiten HIS-HF Absolventenbefragung des Jahrgangs 2005.HIS:Forum Hochschule 14. Hannover : HIS.

Gross, Edward (1968): Universities as Organizations: A Research Approach. In: American Sociological Review. Vol. 33 (4), S. 518-544.

Grove, Wayne A./Wasserman, Tim (2004): The Life-Cycle Pattern of Collegiate GPA: Longitudinal Cohort Analysis and Grade Inflation. In: Journal of Economic Education. Vol. 35 (2), S. 162-174.

Grözinger, Gerd (2015): Einflüsse auf die Notengebung an deutschen Hochschulen. Eine Analyse am Beispiel der Wirtschaftswissenschaften. In: Die Hochschule. Vol. 2/2015, S. 94-114.

Grözinger, Gerd (2017): Einflüsse auf die Notengebung: eine Analyse ausgewählter Fächer auf Basis der Prüfungsstatistik. In: Müller-Benedict, V./Grözinger, G. (Hrsg.): Noten an Deutschlands Hochschulen. Wiesbaden: Springer VS. S. 79-116.

Grüner, Gustav (1971): Die Magisterprüfung in der Bundesrepublik Deutschland. Weinheim : Beltz.

Gump, Steven E. (2007): Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. In: Educational Review Quarterly. Vol. 30 (3), S. 55-68.

Hampe, Asta (1977): Faktoren des Studienverhaltens. Eine vergleichende statistische Untersuchung finanziell geförderter und nicht geförderter Examenskandidaten, Band 1. München : Fink.

Hampe, Asta (1978): Faktoren des Studienverhaltens. Eine vergleichende statistische Untersuchung finanziell geförderter und nicht geförderter Examenskandidaten, Band 2. München : Fink.

Heckhausen, Heinz (1974). Leistung und Chancengleichheit. Motivationsforschung, Band 2. Göttingen : Hogrefe.

Helbig, Marcel (2012): Sind Mädchen besser? Der Wandel geschlechtsspezifischen Bildungserfolgs in Deutschland. Frankfurt a.M. : Campus Verlag.

Helson, Harry (1947): Adaptation-Level as Frame of Reference for Prediction of Psychophysical Data. In: The American Journal of Psychology. Vol. 60 (1). S.1-29.

Helson, Harry (1948): Adaptation-Level as a Basis for a Quantitative Theory of Frames of Reference. In: Psychological Review. Vol. 55(6), S. 297-313.

Hitpass, Josef/Trosien, Jürgen (1987): Leistungsbeurteilung in Hochschulabschlussprüfungen innerhalb von drei Jahrzehnten - Wandel von Prüfungsergebnis und Prüfungserlebnis an deutschen Universitäten. Bad Honnef: Bock.

Hochschulrektorenkonferenz (1991-1993): Übersicht über Studienmöglichkeiten und Zulassungsbeschränkungen für Studienanfänger an den Hochschulen der Bundesrepublik Deutschland. Bonn : HRK.

Hochschulrektorenkonferenz (1993-1996): Studienmöglichkeiten und Zulassungsbeschränkungen für Studienanfänger an den Hochschulen in der Bundesrepublik Deutschland. Bonn : HRK.

Hochschulrektorenkonferenz (1997-2004): Studienangebote deutscher Hochschulen : Studiengänge zum ersten berufsqualifizierenden Abschluss. Bielefeld : Bertelsmann.

Hochschulrektorenkonferenz (2014): Statistische Daten zu Studienangeboten an Hochschulen in Deutschland - Studiengänge, Studierende, Absolventinnen und Absolventen, Wintersemester 2015/2016. Statistiken zur Hochschulpolitik 1/2015. Bonn : HRK.

Horn, Laura/Peter, Katharin/Rooney, Kathryn (2002): Profile of Undergraduates in U.S Postsecondary Institutions: 1999–2000. U.S. Department of Education. National Center for Education Statistics. NCES 2002–168. Washington, DC : 2002.

Hu, Shouping (2005): Beyond Grade Inflation. Grading Problems in Higher Education. San Francisco: Jossey-Bas.

Huber, Ludwig/Portele, Gerhard (1983): Die Hochschullehrer. In: Huber L. (Hrsg.): Ausbildung und Sozialisation in der Hochschule. Enzyklopädie Erziehungswissenschaft. Vol. 10. Stuttgart : Klett-Cotta. S. 193–218.

Huber, Ludwig (1990a): Fachkulturen. Über die Mühen der Verständigung zwischen den Disziplinen. In: Ermert, K. (Hrsg.): Humboldt, High-Tech und High-Culture: was heisst "Hochschulkultur" heute? Loccumer Protokolle 14.Rehburg-Loccum : Evangelische Akademie Loccum. S. 68–99.

Huber, Ludwig (1990b): Disciplinary Cultures and Social Reproduction. In: European Journal of Education. Vol. 25 (3), S. 241-261.

Huber, Ludwig (1990c): Fachkulturen und allgemeine Bildung. In: Lohmann, K. (Hrsg.): Der Beitrag der Unterrichtsfächer zur Allgemeinbildung: Bericht über den 23. Seminartag des Bundesarbeitskreises der Seminar- und Fachleiter e.V. (BAK) vom 2. bis 6. 10. 1989 in der Universität Bielefeld in Zusammenarbeit mit dem Oberstufen-Kolleg der Universität Bielefeld. Mitteilungen des Bundesarbeitskreises der Seminar- und Fachleiter / Bundesarbeitskreis der Seminar- und Fachleiter. Rinteln: Merkur. S. 76–94.

Huber, Ludwig (1991): Sozialisation in der Hochschule. In: Hurrelmann, K./Ulrich, D. (Hrsg.): Neues Handbuch der Sozialisationsforschung. Weinheim : Beltz. S .417-441.

Huber, Ludwig (1992): Neue Lehrkultur-alte Fachkultur. In: Dress, A.W.M. (Hrsg.): Die humane Universität: Bielefeld 1969 - 1992. Festschrift für Karl Peter Grotemeyer. Bielefeld : Westfalen Verlag. S. 95-106.

Huber, Ludwig (1994): Nur allgemeine Studierfähigkeit oder doch allgemeine Bildung? In: Die Deutsche Schule. Vol. 86 (1), S.12-26.

Huber, Michael (2005): Reform in Deutschland. Organisationssoziologische Anmerkungen zur Universitätsreform. In: Soziologie. Vol. 34, S. 391-403.

Hüther, Otto/Krücken, Georg (2016): Hochschulen : Fragestellungen, Ergebnisse und Perspektiven der sozialwissenschaftlichen Hochschulforschung. Wiesbaden : Springer VS.

Ingenkamp, Karlheinz (1995): Die Fragwürdigkeit der Zensurengebung : Texte und Untersuchungsberichte. Weinheim : Beltz.

Ipsen, Detlev (1976): Aspekte der Desorganisation an westdeutschen Hochschulen : eine Zeitbudgetanalyse. In: Ipsen, D./Portele, G. (Hrsg.): Organisation von Forschung und Lehre an westdeutschen Hochschulen. München : Verlag Dokumentation Saur. S. 1-152.

Isely, Paul/Singh, Harinder (2005): Do higher grades lead to favorable student evaluations? In: Journal of Economic Education. Vol. 36 (1), S. 29-42.

Isserstedt, Wolfgang et al. (2010): Die wirtschaftliche und soziale Lage der Studierenden in der Bundesrepublik Deutschland 2009. 19. Sozialerhebung des Deutschen Studentenwerks durchgeführt durch HIS Hochschul-Informations-System. Berlin : Bundesministerium für Bildung und Forschung.

Jacob, Anna, K./Teichler, Ulrich (2011): Der Wandel des Hochschullehrerberufs im internationalen Vergleich. Ergebnisse einer Befragung in den Jahren 2007/08. Bonn : Bundesministerium für Bildung und Forschung.

Jachmann, Michael (2003): Noten oder Berichte? Die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern. Opladen : Leske + Budrich.

Jäger, Reinhold S. (2001): Von der Beobachtung zur Notengebung. Ein Lehrbuch. Landau : Verlag Empirische Pädagogik.

Jewell, R. Todd/McPherson, Michael A. (2012): Instructor-Specific Grade Inflation: Incentives, Gender, and Ethnicity. In: Social Science Quarterly. Vol. 93 (1), S. 95-109.

Jewell, R. Todd/McPherson, Michael A./Tieslau, Margie A. (2013): Whose fault is it? Assigning blame for grade inflation in higher education. In: Applied Economics. Vol. 45, S. 1185–1200.

Jirjahn, Uwe (2007): Welche Faktoren beeinflussen den Erfolg im wirtschaftswissenschaftlichen Studium? In: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung. Vol. 59 (3), S. 286–313.

Johnson, Valen E. (2003): Grade Inflation: A Crisis in College Education. New York : Springer Verlag.

Juola, Arvo E. (1976): Grade Inflation in Higher Education: What Can Or Should We Do? Paper presented at the Annual Meeting of National Council on Measurement in Education in San Francisco, California.

Juola, Arvo E. (1980): Grade Inflation in Higher Education, 1979: Is it Over? Paper presented at the Annual Meeting of National Council on Measurement in Education in Boston, Massachusetts.

Juristische Schulung : JuS - Zeitschrift für Studium und Referendariat. München/Frankfurt a.M. : Beck.

Kadiyala, Koteswara R. (1968): A Transformation Used to Circumvent the Problem of Autocorrelation. In: Econometrica. Vol. 36 (1), S. 93–96.

Kalter, Frank/ Granato, Nadia (2001): Die Persistenz ethnischer Ungleichheit auf dem deutschen Arbeitsmarkt. Diskriminierung oder Unterinvestition in Humankapital? In: Kölner Zeitschrift für Soziologie und Sozialpsychologie. Vol. 53 (3), S. 497-520.

Kalter, Frank/ Granato, Nadia/Kristen, Cornelia (2011): Die strukturelle Assimilation der zweiten Migrantengeneration in Deutschland: Eine Zerlegung gegenwärtiger Trends. In: Becker, Rolf (Hrsg.): Integration durch Bildung. Wiesbaden : VS Verlag für Sozialwissenschaften. S. 257-288.

Kamber, Richard (2008): Understanding Grade Inflation. In: Hunt, L.H. (Hrsg.): Grade Inflation. Academic Standards in Higher Education. New York : State University of New York Press. S. 171-189.

Keil, Wolfgang/Piontkowski, Ursula (1973): Strukturen und Prozesse im Hochschulunterricht. Weinheim : Beltz.

Kezim, Boualem/Pariseau, Susan E., & Quinn, Frances (2005): Is Grade Inflation Related to Faculty Status? In: Journal of Education for Business. Vol. 80 (6), S. 358-363.

Kirp, David L. (2003): Shakespeare, Einstein, and the Bottom Line: The Marketing of Higher Education. Cambridge : Harvard University Press.

Klose, Traugott/Lange, Ernst M. (1981): Diplomprüfungen im Widerstreit. Die Funktion von Hochschulabschlussprüfungen für das Studium und für den Beruf. Symposium am 29. und 30.April 1981. Dokumentationsreihe der Freien Universität Berlin.

Knight, Peter (2006): The local practices of assessment. In: Assessment & Evaluation in Higher Education. Vol. 31 (4), S.435-452.

Köckeis-Stangl, Eva (1972): Über die Desorientierungsfunktion von Schulnoten. Mit empirischen Nachweisen von kontextuellen Determinanten in den Lehrerurteilen über Schüler. Arbeiten aus dem Institut für Erziehungswissenschaft, Band 13/1972. Universität Innsbruck : Institut für Erziehungswissenschaft.

Koedel, Cory (2011): Grading Standards in Education Departments at Universities. In: Education Policy Analysis Archives 19(23). S.1-20.

Köhler, Thomas/Gapski, Jörg (1997): Fachkultur und Lebenswelt Studierender. In: Geiling, H. (Hrsg.): Integration und Ausgrenzung : Hannoversche Forschungen zum gesellschaftlichen Strukturwandel. Hannover : Offizin Verlag. S. 205-234.

Kohn, Alfie (2008): The Dangerous Myth of Grade Inflation. In: Hunt, L.H. (Hrsg.): Grade Inflation. Academic Standards in Higher Education. New York : State University of New York Press. S. 1-12.

Kokkelenberg, Edward C./Dillon, Michael/Christy, Sean M. (2008): The effects of class size on student grades at a public university. In: Economics of Education Review. Vol. 27, S. 221–233.

Kolb, David A. (1981): Learning Styles and Disciplinary Differences. In: Chickering, A.W. et al. (Hrsg.): The Modern American College. San Fransisco : Jossey-Bass.

Kolevzon, Michael S. (1981): Grade inflation in higher education: A comparative study. In: Research in Higher Education. Vol. 15 (3), S. 195-212.

Köller, Olaf (2013): Abitur und Studierfähigkeit. In: Von der Schule zur Hochschule. Analysen, Konzeptionen und Gestaltungsperspektiven des Übergangs. Münster : Waxmann. S. 25-49.

Köller, Olaf/Baumert, Jürgen (2002): Das Abitur - immer noch ein gültiger Indikator für die Studierfähigkeit? In: Aus Politik und Zeitgeschichte Vol. 52, S. 12-19.

Kopp, Botho v./Weiß, Manfred (1995): Der „Arbeitsplatz Universität“ und die Zukunft der Hochschulen. Ergebnisse einer Befragung von Professoren westdeutscher Universitäten. In: Enders, J./Teichler, U. (Hrsg.): Der Hochschullehrerberuf. Aktuelle Studien und ihre hochschulpolitische Diskussion. Neuwied : Luchterhand. S. 105-125.

Krampen, Günter (1984): Welche Funktionen haben Zensuren in der Schule? Eine empirische Untersuchung zu Funktionswahrnehmungen von Lehrern, Lehramtskandidaten und Schülern. In: Zeitschrift für Erziehungswissenschaftliche Forschung. Vol. 18. S. 89-102.

Krais, Beate (1996): The Academic Disciplines: Social Field and Culture. In: Sciulli, D. (Hrsg.): Normative Social Action: Crossnational and Historical Approaches. Greenwich : JAI Press. S. 93-111.

Krautmann, Anthony C./Sander, William (1999): Grades and student evaluations of teachers. In: Economics of Education Review. Vol. 18 (1), S. 59-63.

Krempkow, Rene/König, Karsten/Winter, Jana (2000): Studienführer Sachsen. Dresden : Studentisches Evaluationsbüro Sachsen (SES).

Krempkow, Rene/König, Karsten/Winter, Jana (2001): Studienführer Sachsen. TU Dresden : Institut für Soziologie.

Krempkow, Rene/König, Karsten (2002): Studienführer Sachsen. TU Dresden : Institut für Soziologie.

Krempkow, Rene/König, Karsten (2003): Studienführer Sachsen 2003. TU Dresden : Institut für Soziologie.

Krempkow, Rene/König, Karsten (2004): Studienführer Sachsen 2004. TU Dresden : Institut für Soziologie.

Krempkow, Rene/Popp, Jacqueline/Rachelski, Dietmar (2005): Dokumentation zum „SZ-Hochschul-TÜV“ 2005. TU Dresden : Institut für Soziologie.

Krempkow, Rene (2006): Studienerfolg, Studienqualität und Studierfähigkeit. Eine Analyse zu Determinanten des Studienerfolgs in 150 sächsischen Studiengängen. In: Die Hochschule. Journal für Wissenschaft und Bildung. Vol. 1/2008, S.91-107.

Kristen, Cornelia (2002): Hauptschule, Realschule oder Gymnasium? Ethnische Unterschiede am ersten Bildungsübergang. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie. Vol. 54 (3), S. 534-552.

Kristen, Cornelia (2003): Ethnische Unterschiede im deutschen Schulsystem. In: Aus Politik und Zeitgeschichte. B21-22, S. 26-32.

Kristen, Cornelia/Granato, Nadia (2007): The educational attainment of the second generation in Germany. In: Ethnicities, Vol. 7(3). S. 343-366.

Kristen, Cornelia/Reimer, David/ Kogan, Irena (2008): Higher education entry of Turkish immigrant youth in Germany. In: International Journal of Comparative Sociology. Vol. 49(2-3), S. 127-151.

Kronig, Winfried (2007): Die systematische Zufälligkeit des Bildungserfolgs. Theoretische Erklärungen und empirische Untersuchungen zur Lernentwicklung und zur Leistungsbewertung in unterschiedlichen Schulklassen. Bern : Haupt.

Kuh, George D./Hu, Shouping (1999): Unraveling the Complexity of the Increase in College Grades From the Mid-1980s to the Mid-1990s. In: Educational Evaluation and Policy Analysis. Vol. 21 (3), S.297-321.

Kuh, George D./Whitt, Elizabeth, J. (1988): The Invisible Tapestry. Culture in American Colleges and Universities. ASHE-EPIC Higher Education Report No. 1.

Kuh, George D. (2003): What We're Learning about Student Engagement From NSSE. In: Change. Vol. 35 (2), S. 24-32.

Kultusministerkonferenz (2011): Die Mobilität der Studienanfänger und Studierenden in Deutschland von 1980 bis 2009. Statistische Veröffentlichungen der Kultusministerkonferenz. Dokumentation Nr. 191.

Kühl, Stefan (2012): Nichtangriffspakte an den Hochschulen. Weswegen die Noten an den Hochschulen immer besser werden. Working Paper 5/2012. Universität Bielefeld.

Kwon, Ik-Whan G./Kendig, Nancy L./Bae, Mueun (1997): Grade Inflation From a Career Counselor's Perspective. In: Journal of Employment Counseling. Vol. 34, S. 50-54.

Ladd, Everett C./Lipset, Seymour, M. (1975): The divided academy : Professors and Politics. New York : McGraw-Hill.

Lanning, Wayne/Perkins, Peggy (1995): Grade inflation: A consideration of additional causes. In: Journal of Instructional Psychology. Vol. 22 (2), S. 163-168.

Lawler, Peter A. (2001): Grade Inflation, Democracy, and the Ivy League. In: Perspectives on Political Science. Vol. 30 (3), S.133-136.

Lepenies, Wolf (1985): Die drei Kulturen. Soziologie zwischen Literatur und Wissenschaft. München : Hanser.

Lessing, Hans-Ulrich (2001): Wilhelm Diltheys "Einleitung in die Geisteswissenschaften". Darmstadt : Wissenschaftliche Buchgesellschaft.

Levine, Arthur/Cureton, Jeanette S. (1998): When Hope and Fear Collide. A Portrait of Today's College Student. San Francisco : Jossey-Bass.

Lewin, Dirk/ Lischka, Irene (2004): Passfähigkeit beim Hochschulzugang als Voraussetzung für Qualität und Effizienz von Hochschulbildung. Institut für Hochschulforschung : Arbeitsberichte 6/2004.

Liebau, Eckart/Huber, Ludwig (1985): Die Kulturen der Fächer. In: Neue Sammlung 25 (3). S.314-339.

Lienert, Gustav A./Raatz, Ulrich (1998): Testaufbau und Testanalyse. Weinheim : Beltz.

Lissmann, Urban (1997): Probleme und Möglichkeiten der Schülerbeurteilung. Landau : Verlag Empirische Pädagogik.

Little, Angela (1990): The role of assessment, re-examined in international context. In: Broadfoot, P./Murphy, R./Torrance, H. (Hrsg.): Changing educational assessment. London : Routledge. S. 9-22.

Lowe, S. Keith/Borstorff, Patricia C./Landry III, Robert J. (2008): An empirical examination of the phenomenon of grade inflation in higher education: a focus of grade divergence between business and other fields of study. In: Academy of Educational Leadership Journal. Vol. 12 (1), S. 15-33.

Luhmann, Niklas (1975): Soziologische Aufklärung 2. Aufsätze zur Theorie der Gesellschaft. Opladen : Westdeutscher Verlag.

Luhmann, Niklas (1992): Die Universität als organisierte Institution. In: Dress, A.W.M. (Hrsg.): Die humane Universität: Bielefeld 1969 - 1992. Festschrift für Karl Peter Grotemeyer. Bielefeld : Westfalen Verlag. S. 54-61.

Lundgreen, Peter/Scheunemann, Jana/Schwibbe, Gudrun (2008): Berufliche Schulen und Hochschulen in der Bundesrepublik Deutschland 1949-2001. Datenhandbuch zur deutschen Bildungsgeschichte, Band 8. Göttingen : Vandenhoeck & Ruprecht.

Lundgreen, Peter/Schwibbe, Gudrun/Schallmann, Jürgen (2009): Das Personal an den Hochschulen in der Bundesrepublik Deutschland 1953-2005. Datenhandbuch zur deutschen Bildungsgeschichte, Band 10. Göttingen : Vandenhoeck & Ruprecht.

Madani, Roya/Melzer, Benjamin/Müller, Magnus (2013): Prognose des Studienerfolges im MBA-Studium an der Universität Potsdam. Potsdam : Working Paper.

Maier-Leibnitz, Heinz/Schneider, Christoph (1991): The Status of Academic Research in the Federal German Republic: A Report on Two Surveys and the Testimony of Individual Scientists. In: Minerva. Vol. 29 (1), S. 27-60.

Maiworm, Friedhelm (1989): Zur Notenvergabe an hessischen Hochschulen im Vergleich zum Bundesdurchschnitt. Arbeitspapiere des Wissenschaftlichen Zentrums für Berufs- und Hochschulforschung an der Gesamthochschule Kassel, Vol. 21.

Mangan, Katherine (2009): Professors Compete for Bonuses Based on Student Evaluations. In: The Chronicle of Higher Education vom 30.01.2009.

Mansfield, Harvey C. (2001): Grade Inflation. It’s Time to Face the Facts. In: The Chronicle of Higher Education vom 06.04.2001.

Marsh, Herbert W./Roche, Lawrence A. (2000): Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? In: Journal of Educational Psychology, Vol. 92 (1), S. 202-227.

Martinek, Daniela (2007):Die Ungewissheit im Lehrberuf: Orientierungsstil, Motivationsstrategie und Bezugsnorm-Orientierung bei Lehrer/innen. Hamburg : Kovač.

Mathies, Charles/Webber, Karen (2009): Inflated or Not? An Examination of Grade Change. In: Enrollment Management Journal: Student Access, Finance, and Success in Higher Education. Vol. 3 (3), S. 10-39.

McGrory, Marita (2017): Notengebung bei den Lehramtsstudiengängen: Was bewirkt der Übergang zu den Bologna-Abschlüssen? In: Müller-Benedict, V./Grözinger, G. (Hrsg.): Noten an Deutschlands Hochschulen. Wiesbaden: Springer VS. S. 171-182.

McKenzie, Richard B. (1975): The Economic Effects of Grade Inflation on Instructor Evaluations: A Theoretical Approach. In: The Journal of Economic Education. Vol. 6 (2), S. 99-105.

Mc Spirit, Stephanie/Jones, Kirk E. (1999): Grade Inflation Rates among Different Ability Students, Controlling for Other Factors. In: Education Policy Analysis Archives. Vol. 7 (30), S.1-16.

Metzger, Christoph/ Nüesch, Charlotte (2004): Fair prüfen. Ein Qualitätsleitfaden für Prüfende an Hochschulen. Hochschuldidaktische Schriften, Band 6. St. Gallen : IWP Institut für Wirtschaftspädagogik.

Middendorff, Elke et al. (2013): Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2012. 20. Sozialerhebung des Deutschen Studentenwerks durchgeführt durch das HIS-Institut für Hochschulforschung. Berlin : Bundesministerium für Bildung und Forschung.

Millman, Jason/Slovacek, Simeon P./Kulick, Edward/Mitchell, Karen J. (1983): Does Grade Inflation affect the Reliability of Grades? In: Research in Higher Education. Vol. 19 (4), S. 423-429.

Mintzberg, Henry (1979): The Structuring of Organizations. A Synthesis of the Research. Englewood Cliffs : Prentice-Hall.

Mitchell, Lee C. (1998): Inflation Isn't the Only Thing Wrong With Grading. In: The Chronicle of Higher Education vom 08.05.1998.

Möller, Christina (2015): Herkunft zählt (fast) immer. Soziale Ungleichheiten unter Universitätsprofessorinnen und -professoren. Weinheim : Beltz.

Moore, Melanie/Trahan, Richard (1998). Tenure status and grading practices. In: Sociological Perspectives. Vol. 41 (4), S. 775–781.

Mosler, Karl/Savine, Alexandre (2004): Studienaufbau und Studienerfolg von Kölner Volks- und Betriebswirten im Grundstudium. Diskussionsbeiträge zur Statistik und Ökonometrie 1, Seminar für Wirtschafts- und Sozialstatistik. Köln : Universität zu Köln.

Mullen, Robert (1995): Indicators of Grade Inflation. Paper presented at the annual Forum of the Association for Institutional Research in Boston, Massachusetts.

Müller, Florian H./Bayer, Christina (2007): Prüfungen: Vorbereitung – Durchführung – Bewertung. In: Hawelka, B./Hammerl, M./Gruber, H. (Hrsg.): Förderung von Kompetenzen in der Hochschullehre. Theoretische Konzepte und ihre Implementation in der Praxis. Kröning : Asanger. S. 223-237.

Müller-Benedict, Volker (2005): Sind Universitätsprüfungen objektiv? Der langfristige historische Zusammenhang zwischen Erfolg in akademischen Prüfungen und Überfüllung der akademischen Berufe. In: Soziologie. Vol. 34 (2), S. 191-208.

Müller-Benedict, Volker (2010): Grenzen leistungsbasierter Auswahlverfahren. In: Zeitschrift für Erziehungswissenschaft. Vol. 13 (3), S. 451-472.

Müller-Benedict, Volker/Gaens, Thomas (2015): Sind Examensnoten vergleichbar? Und was, wenn Noten immer besser werden? Der Versuch eines Tabubruchs. In: Die Hochschule. Vol. 2/2015, S. 79-93.

Müller-Benedict, Volker/Janßen, Jörg/Sander, Tobias (2008): Akademische Karrieren in Preußen und Deutschland 1850-1940. Datenhandbuch zur Deutschen Bildungsgeschichte, Band 6. Göttingen: Vandenhoeck & Ruprecht.

Müller-Benedict, Volker/Tsarouha, Elena (2011): Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen. In: Zeitschrift für Soziologie. Vol.40, S. 288-309.

Multrus, Frank(2004): Fachkulturen. Begriffsbestimmung, Herleitung und Analysen: Eine empirische Untersuchung über Studierende deutscher Hochschulen. Dissertation, Universität Konstanz.

Multrus, Frank (2008): Studiensituation und studentische Orientierungen. 10. Studierendensurvey an Universitäten und Fachhochschulen, Langfassung. Berlin : Bundesministerium für Bildung und Forschung.

Mulvenon, Sean/Ferritor, Dan (2005): Grade Inflation in Higher Education. Isolated or Systemic? In: International Journal of Learning. Vol. 12 (6), S.55-61.

Murray, Harry G. (2005): Student evaluation of teaching: Has it made a difference? Paper presented at the Annual Meeting of the Society for Teaching and Learning in Higher Education im Juni 2005. Charlottetown, Prince Edward Island, Canada.

Musselin, Christine (2007): Are Universities specific organisations? In: Krücken G. et al. (Hrsg): Towards a Multiversity? Universities between Global Trends and national Traditions. Bielefeld : Transcript Verlag. S .63-84.

Nath, Axel/Dartenne, Corinna M./Oelerich, Carina (2004): Der historische Pygmalioneffekt der Lehrergeneration im Bildungswachstum von 1884 bis 1993. In: Zeitschrift für Pädagogik. Vol. 50 (4), S. 539-564.

Neumann, Marko/Nagy, Gabriel/Trautwein, Ulrich/ Lüdtke, Oliver (2009): Vergleichbarkeit von Abiturleistungen. Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. In: Zeitschrift für Erziehungswissenschaft. Vol. 12, S. 691–714.

Nierobisch, Kira (2010): Studium, Übergang und Beruf: Unterschiedliche Gestaltungsformen von Pädagog/innen und Mediziner/innen. In: Felden, H.v./Schiener, J. (Hrsg.): Transitionen - Übergänge vom Studium in den Beruf. Zur Verbindung von qualitativer und quantitativer Forschung. Wiesbaden : VS Verlag für Sozialwissenschaften. S. 106-155.

Noelle-Neumann, Elisabeth (1980): Die Arbeitssituation der Professoren : Referat auf dem 30. Hochschulverbandstag 1980 in Hannover. Mitteilungen des Hochschulverbandes. Bonn : Deutscher Hochschulverband. Vol. 28 (3), S. 143-148.

Novy, Diane M./Swank, Paul R./Kopel, Kenneth F. (1996): Psychometrics of Oral Examinations for Psychology Licensure: The Texas Examination as an Example. In: Professional Psychology: Research and Practice. Vol. 27 (4), S. 14-17.

Oehler, Christoph (1986): Zum Rollenwandel der Hochschullehrer. In: Neusel, Aylâ/Teichler, Ulrich (Hrsg.): Hochschulentwicklung seit den sechziger Jahren. Kontinuität - Umbrüche - Dynamik. Weinheim : Beltz. S.257-308.

Oehler, Christoph et al. (1976): Studienplanung und Organisation der Lehre : Ergebnisse einer empirischen Untersuchung in den Hochschulregionen Frankfurt und Darmstadt. Hochschulplanung, Band 25. München : Verlag Dokumentation Saur.

Oehler, Christoph et al. (1978): Organisation und Reform des Studiums : Eine Hochschullehrerbefragung. Hochschulplanung, Band 29. München : Verlag Dokumentation Saur.

Ogilvie, Kristie/Jelavic, Matthew (2013): Grade inflation in the US higher educational environment: a faculty perception study. In: International Journal of Management in Education. Vol. 7 (4), S. 406-416.

Oleinik, Anton (2009): Does education corrupt? Theories of grade inflation. In: Educational Research Review. Vol. 4, S.156-164.

Olsen, Danny R. (1997): Grade Inflation: Reality or Myth? Student Preparation Level vs. Grades at Brigham Young University. Paper presented at the Annual Forum of the Association for Institutional Research in Orlando, Florida.

Ostrovsky, Michael/ Schwarz, Michael (2003): Equilibrium Information Disclosure: Grade Inflation and Unraveling. Harvard Institute of Economic Research Working Papers 01/2003.

Ottwaska, Gertrud (1971): Studienbedingungen, Prüfungsleistungen, Berufserfolg : eine Untersuchung zur Effektivität des Studiums der Wirtschaftswissenschaften an der Universität Mannheim in den Jahren 1958 und 1966. Bielefeld : Bertelsmann.

Parmentier, Klaus/Schade, Hans-Joachim/Schreyer, Franziska (1998): Akademiker/innen - Studium und Arbeitsmarkt. Begleitheft zur Serie. In: Materialien aus der Arbeitsmarkt- und Berufsforschung. Vol. 1.0, S. 1-26.

Parsons, Talcott/Platt, Gerald M. (1973): The American University. Cambridge: University Press.

Pascarella, Ernest T./Terenzini, Patrick T. (2005): How College Affects Students. Volume 2. A Third Decade of Research. San Francisco : Jossey-Bass.

Pattison, Evangeleen/Grodsky, Eric/Muller, Chandra (2013): Is the Sky Falling? Grade Inflation and the Signaling Power of Grades. In: Educational Researcher. Vol. 42 (5), S.259-265.

Pedersen, Daniel (1997): When an A is average. In: Newsweek vom 03.03.1997.

Popov, Sergey V./Bernhardt, Dan (2010): University Competition, Grading Standards and Grade Inflation. MPRA Paper No. 26461.

Portele, Gerhard (1975): Sozialisation in der Hochschule - Vorschläge für ein Forschungsprogramm und einige fachspezifische Ergebnisse. In: Bargel, T. et al. (Hrsg.):Sozialisation in der Hochschule. Beiträge für eine Auseinandersetzung zwischen Hochschuldidaktik und Sozialisationsforschung. Blickpunkt Hochschuldidaktik, Band 37. Hamburg : Arbeitsgemeinschaft für Hochschuldidaktik. S. 96-110.

Portele, Gerhard (1976): Organisation von Lehrveranstaltungen und Lernmotivation : offizielle Lehrveranstaltungen und studentische Arbeitsgemeinschaften. In: Ipsen, D./Portele, G. (Hrsg.): Organisation von Forschung und Lehre an westdeutschen Hochschulen. München : Verlag Dokumentation Saur. S. 155-285.

Pospeschill, Markus (2010): Testtheorie, Testkonstruktion, Testevaluation. Mit 77 Fragen zur Wiederholung. München : Reinhardt.

Potter, William/Nyman, Melvin A./Klumpp, Karen S. (2001): Be careful what you wish for: Analysis of grading trends at a small liberal arts college. In: College and University: The Journal of the American Association of Collegiate Registrars. Vol. 76 (4), S. 9-14.

Prahl, Hans-Werner (1974): Gesellschaftliche Funktionen von akademischen Abschlussprüfungen und Graden : sozialhistorische und ideologiekritische Untersuchungen zur akademischen Initiationskultur. Kiel : Universität Kiel.

Prahl, Hans-Werner (1979): Hochschulprüfungen zwischen Rationalisierung und Ritualisierung - Bemerkungen zum Verhältnis von Prüfungen und Gesellschaftsstruktur. In: Gruppendynamik. Forschung und Praxis. Vol. 10 (4). S. 211-222.

Prahl, Hans-Werner (1983): Prüfungen. In: Huber, L. (Hrsg): Ausbildung und Sozialisation in der Hochschule. Stuttgart : Klett. S. 438-450

Prais, Sigbert J./Winsten, Christopher B. (1954): Trend Estimators and Serial Correlation. Cowles Commission Discussion Paper No. 383 (Chicago).

Prather, James E./Smith, Glynton/Kodras, Janet E. (1979): A Longitudinal Study of Grades in 144 Undergraduate Courses. In: Research in Higher Education. Vol. 10 (1), S.11-24.

Pressman, Steven (2007): The Economics of Grade Inflation. In: Challenge. Vol. 50 (5), S. 93-102.

Preuss, Roland (2012): Zu gute Noten an deutschen Hochschulen. In: Süddeutsche Zeitung vom 10./11.11.2012.

Ramm, Michael/Bargel, Tino (2005): Frauen im Studium. Bonn : Bundesministerium für Bildung und Forschung.

Ramm, Michael/Multrus, Frank/Bargel,Tino (2011): Studiensituation und studentische Orientierungen. 11. Studierendensurvey an Universitäten und Fachhochschulen, Langfassung. Berlin : Bundesministerium für Bildung und Forschung.

Ratzki, Anne (2003): Leistung bewerten. Was Noten leisten - und was nicht. In: Lernende Schule. Vol. 6 (21). S. 4-7.

Rauschenberger, Hans (1999): Umgang mit Schulzensuren. Funktionen-Entwicklungen-Praxis. In: Ebd. (Hrsg.): Leistung und Kontrolle: Die Entwicklung von Zensurengebung und Leistungsmessung in der Schule. Erziehung im Wandel, Band 4. Weinheim : Juventa. S.11-100.

Reiss, Veronika (1975): Die theoretischen Naturwissenschaften als Sozialisationsumwelt für Studenten. In: Bargel, T. et al. (Hrsg.): Sozialisation in der Hochschule. Beiträge für eine Auseinandersetzung zwischen Hochschuldidaktik und Sozialisationsforschung. Blickpunkt Hochschuldidaktik, Band 37. Hamburg : Arbeitsgemeinschaft für Hochschuldidaktik. S. 214-229.

Reisz, Robert, D./Stock, Manfred (2013): Hochschulexpansion, Wandel der Fächerproportionen und Akademikerarbeitslosigkeit in Deutschland. In: Zeitschrift für Erziehungswissenschaft. Vol. 16 (1), S. 137-156.

Rheinberg, Falko (1980): Leistungsbewertung und Lernmotivation. Göttingen : Hogrefe.

Rheinberg, Falko (2002): Bezugsnormen und schulische Leistungsbewertung. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim : Beltz. S. 59-71.

Rogers, Bruce G. (1983): A Time Series Approach to the Longitudinal Study of Undergraduate Grades. Paper presented at the Annual Meeting of the National Council on Measurement in Education in Montreal, Quebec.

Rojstaczer, Stuart (2002/2016): www.gradeinflation.com

Rojstaczer, Stuart/Healy, Christopher (2012): Where A is ordinary: The evolution of American college and university grading, 1940–2009. In: Teachers College Record. Vol. 114 (7), S. 1-23.

Rosovsky, Henry/Hartley, Matthew (2002): Evaluation and the Academy: Are We Doing the Right Thing? Grade Inflation and Letters of Recommendation. Cambridge : American Academy of Arts and Sciences.

Ruscio, Kenneth, P. (1987): Many Sectors, Many Professions. In: Clark, B.R. (Hrsg.): The Academic Profession. National, Disciplinary and Institutional Settings. Berkeley : University of California Press. S. 331-368.

Ruprecht-Karls-Universität Heidelberg (2015): Studierendenstatistik Wintersemester 2014/2015. Heidelberg : Zentrale Universitätsverwaltung Dez. 2 – Studium und Lehre.

Rush, Bonnie R./Elmore, Ronnie G./Sanderson, Michael W. (2009): Grade Inflation at a North American College of Veterinary Medicine: 1985-2006. In: Journal of Veterinary Medical Education. Vol. 36 (1), S. 107-113.

Sabot, Richard/Wakeman-Linn, John (1991): Grade Inflation and Course Choice. In: Journal of Economic Perspectives. Vol. 5 (1), S. 159-170.

Sacher, Werner (1996): Prüfen - beurteilen - benoten : Grundlagen, Hilfen und Denkanstöße für alle Schularten. Bad Heilbrunn : Klinkhardt.

Sadler, Royce, D. (2005): Interpretations of criteria-based assessment and grading in higher education. In: Assessment & Evaluation in Higher Education. Vol. 30 (2), S.175-194.

Salzwedel, Jürgen (1996): Prüfungen. In: Flämig, C. et al. (Hrsg.): Handbuch des Wissenschaftsrechts, Band 1. Berlin : Springer. S. 731-750.

Schaeper, Hildegard (1995): Zur Arbeitssituation von Lehrenden an westdeutschen Universitäten. Ergebnisse einer empirischen Untersuchung in fünf ausgewählten Disziplinen. In: Enders, J./Teichler, U. (Hrsg.): Der Hochschullehrerberuf. Aktuelle Studien und ihre hochschulpolitische Diskussion. Neuwied : Luchterhand. S. 127-153.

Schaeper, Hildegard (1997): Lehrkulturen, Lehrhabitus und die Struktur der Universität. Eine empirische Untersuchung fach- und geschlechtsspezifischer Lehrkulturen. Blickpunkt Hochschuldidaktik, Band 100. Weinheim : Deutscher Studien Verlag.

Schimank, Uwe (1992): Forschungsbedingungen der Professoren an den westdeutschen Hochschulen. Daten aus einer Befragung im Wintersemester 1990/91. MPIFG Discussion Paper Vol.92 (2).

Schimank, Uwe (2013): Gesellschaft. Bielefeld : Transcript.

Schlicht, Uwe (2012): Kuschelnoten: Eine Eins für alle. In: Tagesspiegel vom 13.11.2012.

Schneider, Geoffrey E. (2013): Student Evaluations, Grade Inflation and Pluralistic Teaching: Moving from Customer Satisfaction to Student Learning and Critical Thinking. In: Forum for Social Economics. Vol. 42 (1), S. 122-135.

Schölling, Markus (2005): Soziale Herkunft, Lebensstil und Studienfachwahl : eine Typologie. Frankfurt a.M. : Peter Lang.

Schulz, Florian/Zehner, Fabian/Schindler, Christoph/Prenzel, Manfred (2014): Prüfen und Lernen im Studium: Erste Schritte zur Untersuchung von Prüfungsanforderungen und Lerntypen. In: Beiträge zur Hochschulforschung. Vol. 36 (2), S. 34-58.

Schwager, Robert (2008): Grade Inflation, Social Background, and Labour Market Matching. ZEW Discussion Paper No. 08-070.

Shay, Suellen (2005): The assessment of complex tasks: a double reading. In: Studies in Higher Education. Vol.30 (6), S.663-679.

Snow, Charles P. (1959): The Two Cultures and the Scientific Revolution. Cambridge : University Press.

Sonner, Brenda S. (2000): A is for “Adjunct”: Examining Grade Inflation in Higher Education. In: Journal of Education for Business. Vol. 76 (1), S.5-8.

Spiewack, Matthias (2003): Noten ohne Wert. In: Zeit vom 20.02.2003.

Staples, Brent (1998): Why Colleges Shower Their Students With A's. In: New York Times vom 08.03.1998.

Statistisches Bundesamt (1959-1976): Bildung und Kultur - Studierende an Hochschulen (Fachserie A, Reihe 10.5). Wiesbaden : Statistisches Bundesamt.

Statistisches Bundesamt (1977-2014): Bildung und Kultur - Studierende an Hochschulen (Fachserie 11, Reihe 4.1). Wiesbaden : Statistisches Bundesamt.

Statistisches Bundesamt (2013): Bildung und Kultur – Finanzen der Hochschulen (Fachserie 11, Reihe 4.5). Wiesbaden : Statistisches Bundesamt.

Statistisches Landesamt Schleswig-Holstein, Forschungsdatenzentrum (2012): Hochschulprüfungsstatistik 1995–2010. Zugang über Arbeitsplatz für Gastwissenschaftler.

Stief, Mahena/Abele, Andrea E. (2002): Berufsstart - Sozialwissenschaftler und Sozialwissenschaftlerinnen im Vergleich mit anderen Fächern : Befunde aus einer Langzeitstudie. In: Sozialwissenschaften und Berufspraxis. Vol. 25 (1/2), S. 85-98.

Stieler, Jona.F. (2011) Validität summativer Prüfungen. Überlegungen zur Gestaltung von Klausuren. Bielefeld : Janus.

Stier, Winfried (2001): Methoden der Zeitreihenanalyse. Berlin : Springer.

Stone, John E. (1995): Inflated Grades, Inflated Enrollment, and Inflated Budgets: An Analysis and Call for Review at the State Level. In: Education Policy Analysis Archives. Vol. 3 (11), S. 1-30.

Strenta, A. Christopher/Elliot, Rogers (1987): Differential Grading Standards Revisited. In: Journal of Educational Measurement. Vol. 24 (4), S. 281-291.

Struck, Olaf (2001): Gatekeeping zwischen Individuum, Organisation und Institution. In: Leisering, L./Müller, R./Schumann, K.F. (Hrsg.): Institutionen und Lebensläufe im Wandel. Weinheim : Juventa. S. 29-54.

Studienstiftung des deutschen Volkes e.V. (2015): Jahresbericht 2014.

Südkamp, Anna/Möller, Jens (2009): Referenzgruppeneffekte im Klassenraum. Direkte und indirekte Einschätzungen von Schülerleistungen. In: Zeitschrift für pädagogische Psychologie. Vol. 23 (3-4), S.161-174.

Suslow, Sidney (1976): A Report on an Interinstitutional Survey of Undergraduate Scholastic Grading 1960s to 1970s. California University, Berkeley : Office of Institutional Research.

Teichler Ulrich et al. (1987): Hochschule – Studium – Berufsvorstellungen. Eine empirische Untersuchung zur Vielfalt von Hochschule und deren Auswirkungen. Studien zu Bildung und Wissenschaft, Band 50. Bad Honnef : Bundesministerium für Bildung und Wissenschaft.

Thome, Helmut (2005): Zeitreihenanalyse : Eine Einführung für Sozialwissenschaftler und Historiker. München : Oldenbourg.

Titze, Hartmut (1990): Der Akademikerzyklus: Historische Untersuchungen über die Wiederkehr von Überfüllung und Mangel in akademischen Karrieren. Göttingen : Vandenhoeck & Ruprecht.

Towfigh, Emanuel/Traxler, Christian/Glöckner, Andreas (2014): Zur Benotung in der Examensvorbereitung und im ersten Examen. Eine empirische Analyse. In: Zeitschrift für Didaktik der Rechtswissenschaft. Vol. 1, S. 8-27.

Trapmann, Sabrina/Hell, Benedikt/Weigand, Sonja/Schuler, Heinz (2007): Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse. In: Zeitschrift für Pädagogische Psychologie. Vol. 21 (1), S. 11–27.

Trautwein, Ulrich/Baeriswyl, Franz (2007): Wenn leistungsstarke Klassenkameraden ein Nachteil sind: Referenzgruppeneffekte bei Übergangsentscheidungen. In: Zeitschrift für Pädagogische Psychologie. Vol. 21, S. 119–133.

Unispiegel (2007): Einsen für alle: Kuschelnoten, Kuhhandel, Kumpanei. In: Unispiegel vom 17.01.2007.

Van den Bussche, Hendrik/Wegscheider, Karl/Zimmermann, Thomas (2006): Der Ausbildungserfolg im Vergleich (II). In: Deutsches Ärzteblatt. Vol. 103 (34-35), S.2225-2228.

Von Dietrich, Walther (1984): Abschlussprüfungen mit Noten an Hochschulen in den Prüfungsjahren 1980 und 1982. In: Statistische Rundschau für das Land Nordrhein-Westfalen. Vol. 36 (1), S. 650-661 und S. 691-705.

Von Holdt, Ulrike/Schneider, H./Wagner, Bernardo (2006): Analyse von Studienverläufen und Studienabbrüchen in den Bachelorstudiengängen Informatik an der Leibniz Universität Hannover, HDI 2006: Hochschuldidaktik der Informatik. In: GI-Edition Lecture Notes in Informatics - Proceedings, 7.-8.12.2006, München.

Warning, Susanne/Welzel, Peter (2005): A Note on Grade Inflation and University Competition. Paper based on presentations given at the 2005 meeting of the Allied Social Sciences Association in Philadelphia and at the 3rd Workshop on Business and Economic Policy in Sion.

Wass, Val/Wakeford, Richard/Neighbour, Roger/Van der Vleuten, Cees (2003): Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. In Medical Education. Vol. 37 (2), S. 126-131.

Webler, Wolff-Dietrich (2010): Internationale Vergleichbarkeit von Noten im Hochschulbereich? Problematik der Notenvergabe, Referenzgrößen und der Verwendung der Gaußschen Normalverteilung. In: Qualität in der Wissenschaft. Zeitschrift für Qualitätsentwicklung in Forschung, Studium und Administration. Vol. 4 (1), S. 20-23.

Wenger, Etienne (1998): Communities of practice. Learning, meaning and identity. Cambridge : University Press.

Weick, Karl E. (1976): Educational Organizations as Loosely Coupled Systems. In: Administrative Science Quarterly. Vol. 21, S. 1-19.

Weigand, Dominik (2012): Die Macht der Fachkultur. Eine vergleichende Analyse fachspezifischer Studienstrukturen. Marburg : Tectum.

Weinberg, Bruce A./Hashimoto, Masanori/Fleisher, Belton M. (2009): Evaluating Teaching in Higher Education. In: Journal of Economic Education. Vol. 40 (3), S. 227-261.

Weiss, Robert M./ Rasmussen, Glen R. (1960): Grading Practices in Undergraduate Education Courses: Are the Standards Too Low? In: The Journal of Higher Education. Vol. 31 (3), S. 143-149.

Westdeutsche Rektorenkonferenz (1969-1971): Übersicht über die Zulassungsbeschränkungen für deutsche Studierende an den Hochschulen. Bonn : WRK.

Westdeutsche Rektorenkonferenz (1972-1980): Übersicht über Studienmöglichkeiten und Zulassungsbeschränkungen für deutsche Studienanfänger an den Hochschulen der Bundesrepublik Deutschland. Bonn : WRK.

Westdeutsche Rektorenkonferenz (1981-1990): Übersicht über Studienmöglichkeiten und Zulassungsbeschränkungen für Studienanfänger an den Hochschulen der Bundesrepublik Deutschland. Bonn : WRK.

Weyer, Marc (2013): Rechnungslegungsgestützte Leistungsmessung von Hochschulen in NRW. Duisburg : Universitätsbibliothek Duisburg-Essen.

Whitley, Richard (1984): The intellectual and social organization of the sciences. Oxford : Clarendon.

Wilcke, Bernd-Achim (1976): Studienmotivation und Studienverhalten. Göttingen : Verlag für Psychologie Hogrefe.

Wilson, Bradford P. (1999): The Phenomenon of Grade Inflation in Higher Education. In: National Forum. Vol. 79 (4), S. 38-41.

Windolf, Paul (1992): Fachkultur und Studienfachwahl. Ergebnisse einer Befragung von Studienanfängern. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie. Vol. 44 (1), S. 76-98.

Winteler, Adi (1981): The academic department as environment for teaching and learning. In: Higher Education. Vol. 10 (1), S. 25-35.

Winzer, Margret (2002): Grade Inflation. An Appraisal of the Research. Technical Report, University of Lethbridge Faculty of Education.

Wissenschaftsrat (1988): Grunddaten zum Personalbestand der Hochschulen 1983. Drucksache 8000-88.

Wissenschaftsrat (1995): Grunddaten zum Personalbestand der Hochschulen 1989. Drucksache 1816-94.

Wissenschaftsrat (1998a): Empfehlungen zur Differenzierung des Studiums durch Teilzeitstudienmöglichkeiten. Drucksache 3535-98.

Wissenschaftsrat (1998b): Grunddaten zum Personalbestand der Hochschulen 1995. Drucksache 3721-98.

Wissenschaftsrat (1999): Stellungnahme zum Verhältnis von Hochschulausbildung und Beschäftigungssystem. Drucksache 4099-99.

Wissenschaftsrat (2003): Prüfungsnoten an Hochschulen 1996, 1998 und 2000 nach ausgewählten Studienbereichen und Studienfächern – Arbeitsbericht. Hrsg. v. d. Geschäftsstelle des Wissenschaftsrats, Drucksache 5536–03.

Wissenschaftsrat (2007): Prüfungsnoten im Prüfungsjahr 2005 an Universitäten (einschließlich KH, PH, TH) sowie an Fachhochschulen (einschließlich Verwaltungsfachhochschulen) nach ausgewählten Studienbereichen und Studienfächern – Arbeitsbericht . Hrsg. v. d. Geschäftsstelle des Wissenschaftsrats, Drucksache 7769–07.

Wissenschaftsrat (2012): Prüfungsnoten an Hochschulen im Prüfungsjahr 2010 - Arbeitsbericht mit einem wissenschaftspolitischen Kommentar des Wissenschaftsrates. Hrsg. v. d. Geschäftsstelle des Wissenschaftsrats, Drucksache 2627-12.

Wittenberg, Reinhard (2005):Einflussgrößen auf Studienerfolg, Stellensuche und Einkommen von Sozialwissenschaftlern : ausgewählte Ergebnisse der vierten Umfrage unter Absolventen des Studiengangs Sozialwissenschaften an der Universität Erlangen-Nürnberg. In: Sozialwissenschaften und Berufspraxis . Vol. 28 (2), S. 250-269.

Wolf, Christof/Best, Henning (2010): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden : VS Verlag für Sozialwissenshaften.

Wongsurawat, Winai (2009): Does grade inflation affect the credibility of grades? Evidence from US law school admissions. In: Education Economics. Vol. 17 (4), S.523-534.

Wooldridge, Jeffrey M. (2009): Introductory Econometrics : A modern approach. Mason, Ohio : South-Western, Cengage Learning.

Yang, Huanxing/Yip, Chun Seng (2003): An Economic Theory of Grade Inflation. Working Paper, University of Pennsylvania.

Yorke, Mantz et al. (1996): Module mark distributions in eight subject areas and some issues they raise. In: Jackson, N. (Hrsg.): Modular higher education in the UK. London: Higher Education Quality Council. S. 105-107.

Yorke, Mantz (2008): Grading Student Achievement in Higher Education. Signals and Shortcomings. Abingdon : Routledge.

Yorke, Mantz (2009): Honours degree classification: What we can and cannot tell from the statistics. Gloucester : Quality Assurance Agency for Higher Education.

Yorke, Mantz (2011): Summative assessment: dealing with the 'measurement fallacy'. In: Studies in Higher Education. Vol. 36 (3), S. 251-273.

Zentralstelle für die Vergabe von Studienplätzen (2005): Auswahl- und Verteilungsgrenzen in bundesweit zulassungsbeschränkten Studiengängen zum Wintersemester 2005/2006. Studiengang Biologie. Dortmund : Informations- und Pressestelle der ZVS.

Zentralstelle für die Vergabe von Studienplätzen (2005): Auswahl- und Verteilungsgrenzen in bundesweit zulassungsbeschränkten Studiengängen zum Wintersemester 2005/2006. Studiengang Medizin. Dortmund : Informations- und Pressestelle der ZVS.

Zentralstelle für die Vergabe von Studienplätzen (2005): Auswahl- und Verteilungsgrenzen in bundesweit zulassungsbeschränkten Studiengängen zum Wintersemester 2005/2006. Studiengang Psychologie. Dortmund : Informations- und Pressestelle der ZVS.

Ziegenspeck, Jörg (1999): Handbuch Zensur und Zeugnis in der Schule: Historischer Rückblick, allgemeine Problematik, empirische Befunde und bildungspolitische Implikationen. Bad Heilbrunn : Klinkhardt.

Zirkel, Perry A. (1999): Grade Inflation: A Leadership Opportunity for Schools of Education? In: Teachers College Record. Vol. 101 (2), S.247-260.

Zubrickas, Robertas (2010): Optimal Grading. Working Paper No. 487. Institute for Empirical Research in Economics - University of Zürich Working Paper Series.

ZVS Informations- und Pressestelle (2005): Auswahl- und Verteilungsgrenzen in bundesweit zulassungsbeschränkten Studiengängen zum Wintersemester 2005/2006.

Anhang

Tabelle A1: Fallzahlen der Stichprobe in den einzelnen Studiengängen seit 1950

Jahr	Biologie	Chemie	Mathematik	Psychologie	BWL	VWL	Maschinen-bau	Soziologie Magister	Germanistik Magister	Mathematik Lehramt	Deutsch Lehramt	Jura
1950		61 (3)				94 (2)						
1951		46 (3)				124 (2)						
1952		58 (3)				113 (2)						
1953		43 (3)	18 (2)			213 (3)						
1954		64 (3)	21 (2)			185 (3)						
1955		84 (3)	19 (2)			205 (4)						
1956		70 (3)	20 (2)			233 (4)						
1957		83 (3)	20 (2)		67 (2)	246 (4)						
1958		76 (3)	10 (2)		70 (2)	226 (4)						
1959		128 (4)	11 (2)	13 (2)	135 (2)	146 (4)						3028 (10)
1960		180 (5)	24 (3)	23 (3)	146 (2)	202 (5)						3400 (10)
1961		145 (5)	20 (3)	18 (3)	131 (2)	166 (5)				38 (2)		3283 (10)
1962		167 (5)	21 (3)	25 (3)	164 (2)	178 (5)				44 (2)		3305 (10)
1963		183 (5)	35 (4)	26 (3)	224 (3)	176 (6)				64 (4)	166 (3)	3150 (10)
1964		194 (5)	41 (5)	47 (3)	257 (4)	221 (6)			2 (2)	93 (4)	188 (3)	2792 (10)
1965		197 (5)	46 (5)	47 (3)	262 (4)	277 (6)			4 (2)	111 (4)	231 (3)	2698 (10)
1966		173 (5)	74 (6)	45 (3)	425 (4)	367 (6)			6 (2)	103 (4)	211 (3)	2850 (10)
1967	13 (2)	216 (5)	114 (6)	79 (4)	433 (4)	370 (6)		10 (2)	6 (2)	106 (4)	227 (3)	3089 (10)
1968	13 (2)	196 (5)	106 (6)	137 (4)	538 (4)	446 (6)		8 (2)	8 (2)	153 (4)	243 (3)	3465 (10)
1969	23 (2)	184 (5)	135 (6)	128 (4)	512 (4)	460 (6)		14 (3)	13 (3)	107 (4)	282 (3)	4284 (10)
1970	33 (3)	217 (7)	109 (6)	116 (4)	500 (4)	384 (6)		21 (4)	25 (6)	87 (4)	218 (3)	3712 (10)
1971	45 (4)	285 (7)	140 (7)	145 (5)	442 (4)	387 (6)		18 (4)	57 (6)	101 (4)	256 (3)	3532 (10)
1972	104 (4)	299 (7)	157 (7)	172 (6)	451 (4)	305 (6)	173 (2)	28 (4)	48 (6)	124 (4)	290 (4)	4359 (10)
1973	100 (4)	311 (7)	196 (7)	205 (6)	519 (4)	256 (6)	181 (2)	40 (4)	55 (6)	213 (5)	440 (5)	5132 (10)
1974	136 (4)	308 (7)	269 (7)	269 (6)	500 (4)	319 (6)	179 (2)	16 (4)	47 (6)	238 (5)	699 (5)	4887 (10)
1975	181 (4)	317 (7)	272 (7)	330 (6)	534 (4)	336 (6)	206 (2)	44 (4)	49 (6)	232 (5)	705 (5)	4326 (10)
1976	196 (5)	296 (7)	233 (7)	368 (6)	523 (4)	351 (6)	232 (2)	32 (4)	79 (6)	207 (5)	623 (5)	3496 (10)
1977	173 (5)	247 (7)	252 (7)	416 (6)	610 (4)	370 (6)	214 (2)	44 (4)	87 (6)	303 (5)	808 (5)	3857 (10)
1978	174 (5)	306 (7)	221 (7)	346 (6)	583 (4)	460 (6)	271 (2)	63 (4)	108 (6)	406 (5)	819 (5)	4324 (10)
1979	187 (5)	266 (7)	237 (7)	359 (6)	571 (4)	434 (6)	277 (2)	46 (4)	120 (6)	367 (5)	974 (5)	4930 (10)
1980	180 (5)	310 (7)	230 (7)	583 (6)	491 (4)	316 (6)	325 (2)	46 (4)	174 (6)	419 (5)	918 (5)	5580 (10)
1981	205 (6)	335 (7)	202 (7)	519 (6)	411 (4)	253 (6)	356 (2)	41 (4)	191 (6)	413 (5)	879 (5)	6158 (10)
1982	291 (7)	330 (7)	195 (7)	465 (6)	571 (3)	307 (6)	297 (2)	31 (4)	172 (6)	362 (5)	854 (5)	5592 (10)
1983	321 (7)	348 (7)	210 (7)	471 (6)	536 (3)	281 (6)	313 (2)	35 (4)	214 (6)	224 (5)	504 (5)	5535 (10)
1984	412 (7)	317 (7)	164 (7)	501 (6)	515 (4)	297 (6)	346 (2)	45 (4)	242 (6)	159 (5)	524 (5)	5854 (10)
1985	444 (7)	372 (7)	161 (7)	494 (6)	678 (4)	219 (6)	372 (2)	38 (4)	267 (6)	164 (5)	552 (5)	6015 (10)
1986	447 (7)	315 (7)	171 (7)	460 (6)	771 (4)	227 (6)	413 (2)	56 (4)	300 (6)	146 (5)	507 (5)	7082 (10)
1987	587 (7)	382 (7)	215 (7)	538 (6)	869 (4)	249 (6)	508 (2)	66 (4)	342 (6)	92 (5)	412 (5)	6951 (10)
1988	619 (7)	483 (7)	235 (7)	435 (6)	961 (4)	265 (6)	498 (2)	84 (4)	391 (6)	68 (5)	342 (5)	7927 (10)
1989	658 (7)	475 (7)	258 (7)	438 (6)	1107 (4)	280 (6)	578 (2)	76 (4)	398 (6)	58 (5)	298 (5)	8020 (10)
1990	725 (7)	484 (7)	254 (7)	450 (6)	1297 (4)	271 (6)	668 (2)	65 (4)	411 (6)	50 (5)	252 (5)	8175 (11)
1991	792 (7)	513 (7)	276 (7)	444 (6)	1455 (4)	288 (6)	643 (2)	68 (4)	365 (6)	58 (5)	250 (5)	7508 (11)
1992	810 (7)	498 (7)	238 (7)	474 (6)	1532 (4)	359 (6)	657 (2)	53 (4)	390 (6)	65 (5)	209 (5)	8411 (11)
1993	807 (7)	529 (7)	289 (7)	563 (6)	1695 (4)	408 (6)	711 (2)	66 (4)	389 (6)	72 (5)	286 (5)	9752 (11)
1994	844 (7)	556 (7)	323 (7)	577 (6)	1644 (4)	496 (6)	727 (2)	76 (4)	417 (6)	141 (5)	350 (5)	10017 (11)
1995	826 (7)	527 (7)	296 (7)	479 (6)	1515 (4)	558 (6)	730 (2)	70 (4)	390 (6)	206 (5)	424 (5)	10812 (11)
1996	677 (7)	510 (7)	318 (7)	450 (6)	1369 (4)	533 (6)	751 (2)	58 (4)	411 (6)	290 (5)	466 (5)	11424 (11)
1997	670 (7)	432 (7)	310 (7)	516 (6)	1338 (4)	540 (6)	615 (2)	66 (4)	383 (6)	308 (5)	550 (5)	11124 (11)
1998	550 (7)	330 (7)	207 (7)	603 (6)	1134 (4)	428 (6)	560 (2)	57 (4)	368 (6)	27 (3)	88 (3)	10709 (11)
1999	543 (7)	215 (7)	187 (7)	448 (6)	1001 (4)	323 (6)	420 (2)	53 (4)	240 (6)	30 (3)	103 (3)	10605 (11)
2000	558 (7)	214 (7)	130 (7)	413 (6)	902 (4)	294 (6)	317 (2)	62 (4)	289 (6)	24 (3)	90 (3)	10371 (11)
2001	545 (7)	198 (7)	144 (7)	417 (6)	996 (4)	258 (6)	278 (2)	49 (4)	300 (6)	21 (3)	78 (3)	9624 (11)
2002	564 (7)	194 (7)	142 (7)	451 (6)	969 (4)	231 (6)	263 (2)	58 (4)	298 (6)	21 (3)	102 (3)	9466 (11)
2003	608 (7)	189 (7)	156 (7)	693 (6)	878 (4)	232 (6)	291 (2)	116 (4)	364 (6)	23 (3)	93 (3)	8357 (11)
2004	623 (7)	201 (7)	132 (7)	591 (6)	981 (4)	285 (6)	316 (2)	69 (4)	321 (6)	27 (3)	71 (3)	8514 (11)
2005	621 (7)	223 (7)	127 (7)	539 (6)	964 (4)	308 (6)	312 (2)	96 (4)	339 (6)	20 (3)	115 (4)	8053 (11)
2006	617 (7)	270 (7)	157 (7)	592 (6)	964 (4)	354 (6)	383 (2)	103 (4)	380 (6)	18 (3)	222 (4)	8999 (11)
2007	705 (7)	246 (7)	233 (7)	572 (6)	988 (4)	499 (6)	401 (2)	129 (4)	428 (6)	26 (3)	254 (4)	8887 (11)
2008	678 (7)	295 (7)	243 (7)	504 (6)	881 (4)	482 (5)	452 (2)	119 (4)	404 (6)	36 (4)	230 (4)	
2009	542 (6)	322 (7)	210 (7)	496 (6)	820 (4)	472 (5)	472 (2)	129 (4)	363 (6)	41 (3)	228 (4)	
2010	306 (5)	201 (7)	182 (7)	338 (6)	498 (4)	311 (5)	358 (2)	60 (4)	257 (6)	9 (2)	83 (4)	
Total	**19 153**	**16 214**	**9 416**	**18 828**	**39 328**	**18 874**	**16 064**	**2 524**	**10 912**	**6 715**	**17 714**	**313 421**

Tabelle A2: Fehlende und mittels Linearer Interpolation ersetzte Werte in der Stichprobe

	Göttingen	Braunschweig	Karlsruhe	Berlin	Tübingen	Heidelberg	Münster	Saarbrücken
Soziologie					1967*, 1968*	1998*, 1999*, 2002-2006*		
Germanistik	1999*					1998*, 2003*, 2005*	1965***	1974***, 1978***
Psychologie						2002-2005*		
Mathe		2001*, 2004*		1957***		1998*, 2004*, 2005*, 2007*, 2008*		
Biologie				1959***		2003*, 2005*, 2010**	1981*, 1982*,	
Chemie						1998*, 1999*, 2002*, 2003*, 2005*	2003*, 2009**, 2010**	
VWL			2000*, 2008-2010**			1973*, 1975*, 1976*, 2003*, 2005*	2006*	
BWL			1979***					
Lehramt Deutsch			1998-2004**		1998-2010**			
Lehramt Mathe	2000*, 2004*	1991***,2000*, 2009**,2010**	1998-2007**	2001*, 2002*, 2006*,2010**	1998-2010**			

* Es haben Prüfungen stattgefunden, es sind aber keine Informationen vorliegend (bis 1997: keine Archivdaten, ab 1998: keine FDZ Daten), fehlende Werte wurden mittels linearer Interpolation ersetzt

** Es haben Prüfungen stattgefunden, es sind aber keine Informationen vorliegend (bis 1997: keine Archivdaten, ab 1998: keine FDZ Daten), fehlende Werte wurden *nicht* ersetzt

** *Es haben keine Prüfungen stattgefunden, fehlende Werte wurden mittels linearer Interpolation ersetzt

Tabelle A3: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Jura

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-1.704	0.032	-53.63	0.000
Mathematik Lehramt	-1.236	0.032	-38.88	0.000
Chemie	-1.693	0.032	-53.27	0.000
Biologie	-1.917	0.032	-60.32	0.000
Psychologie	-1.840	0.032	-57.92	0.000
Maschinenbau	-1.459	0.032	-45.90	0.000
VWL	-0.856	0.032	-26.94	0.000
BWL	-0.606	0.032	-19.08	0.000
Soziologie Magister	-1.391	0.032	-43.76	0.000
Germanistik Magister	-1.444	0.032	-45.44	0.000
Deutsch Lehramt	-1.173	0.032	-36.91	0.000
Konstante	3.331	0.022	148.25	0.000
				n=432; r^2_{adj}=0.94

Tabelle A4: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Deutsch Lehramt

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-0.531	0.032	-16.73	0.000
Mathematik Lehramt	-0.063	0.032	-1.98	0.049
Chemie	-0.520	0.032	-16.37	0.000
Biologie	-0.744	0.032	-23.42	0.000
Psychologie	-0.668	0.032	-21.02	0.000
Maschinenbau	-0.286	0.032	-9.00	0.000
VWL	0.317	0.032	9.97	0.000
BWL	0.566	0.032	17.83	0.000
Soziologie Magister	-0.218	0.032	-6.86	0.000
Germanistik Magister	-0.271	0.032	-8.54	0.000
Jura	1.173	0.032	36.91	0.000
Konstante	2.158	0.022	96.06	0.000
				n=432; r^2_{adj}=0.94

Tabelle A5: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Germanistik Magister

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-0.260	0.032	-8.19	0.000
Mathematik Lehramt	0.208	0.032	6.56	0.000
Chemie	-0.249	0.032	-7.83	0.000
Biologie	-0.473	0.032	-14.88	0.000
Psychologie	-0.396	0.032	-12.48	0.000
Maschinenbau	-0.015	0.032	-0.46	0.647
VWL	0.588	0.032	18.51	0.000
BWL	0.838	0.032	26.37	0.000
Soziologie Magister	0.053	0.032	1.68	0.093
Deutsch Lehramt	0.271	0.032	8.54	0.000
Jura	1.444	0.032	45.44	0.000
Konstante	1.887	0.022	83.98	0.000
				n=432; r^2_{adj}=0.94

Tabelle A6: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Soziologie Magister

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-0.314	0.032	-9.87	0.000
Mathematik Lehramt	0.155	0.032	4.88	0.000
Chemie	-0.302	0.032	-9.51	0.000
Biologie	-0.526	0.032	-16.56	0.000
Psychologie	-0.450	0.032	-14.16	0.000
Maschinenbau	-0.068	0.032	-2.14	0.033
VWL	0.535	0.032	16.83	0.000
BWL	0.784	0.032	24.68	0.000
Germanistik Magister	-0.053	0.032	-1.68	0.093
Deutsch Lehramt	0.218	0.032	6.86	0.000
Jura	1.391	0.032	43.76	0.000
Konstante	1.940	0.022	86.36	0.000
				n=432; r^2_{adj}=0.94

Tabelle A7: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie BWL

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-1.098	0.032	-34.55	0.000
Mathematik Lehramt	-0.629	0.032	-19.80	0.000
Chemie	-1.086	0.032	-34.19	0.000
Biologie	-1.310	0.032	-41.24	0.000
Psychologie	-1.234	0.032	-38.84	0.000
Maschinenbau	-0.852	0.032	-26.82	0.000
VWL	-0.250	0.032	-7.86	0.000
Soziologie Magister	-0.784	0.032	-24.68	0.000
Germanistik Magister	-0.838	0.032	-26.37	0.000
Deutsch Lehramt	-0.566	0.032	-17.83	0.000
Jura	0.606	0.032	19.08	0.000
Konstante	2.725	0.022	121.27	0.000
				n=432; r^2_{adj}=0.94

Tabelle A8: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie VWL

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-0.848	0.032	-26.67	0.000
Mathematik Lehramt	-0.380	0.032	-11.95	0.000
Chemie	-0.837	0.032	-26.33	0.000
Biologie	-1.061	0.032	-33.39	0.000
Psychologie	-0.985	0.032	-30.99	0.000
Maschinenbau	-0.603	0.032	-18.97	0.000
BWL	0.250	0.032	7.86	0.000
Soziologie Magister	-0.535	0.032	-16.83	0.000
Germanistik Magister	-0.588	0.032	-18.51	0.000
Deutsch Lehramt	-0.317	0.032	-9.97	0.000
Jura	0.856	0.032	26.94	0.000
Konstante	2.475	0.022	110.16	0.000
				n=432; r^2_{adj}=0.94

Tabelle A9: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Maschinenbau

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-0.246	0.032	-7.73	0.000
Mathematik Lehramt	0.223	0.032	7.02	0.000
Chemie	-0.234	0.032	-7.37	0.000
Biologie	-0.458	0.032	-14.42	0.000
Psychologie	-0.382	0.032	-12.02	0.000
VWL	0.603	0.032	18.97	0.000
BWL	0.852	0.032	26.82	0.000
Soziologie Magister	0.068	0.032	2.14	0.033
Germanistik Magister	0.015	0.032	0.46	0.647
Deutsch Lehramt	0.286	0.032	9.00	0.000
Jura	1.459	0.032	45.90	0.000
Konstante	1.872	0.022	83.33	0.000
				n=432; r^2_{adj}=0.94

Tabelle A10: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Psychologie

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	0.136	0.032	4.29	0.000
Mathematik Lehramt	0.605	0.032	19.04	0.000
Chemie	0.148	0.032	4.65	0.000
Biologie	-0.076	0.032	-2.40	0.017
Maschinenbau	0.382	0.032	12.02	0.000
VWL	0.985	0.032	30.99	0.000
BWL	1.234	0.032	38.84	0.000
Soziologie Magister	0.450	0.032	14.16	0.000
Germanistik Magister	0.396	0.032	12.48	0.000
Deutsch Lehramt	0.668	0.032	21.02	0.000
Jura	1.840	0.032	57.92	0.000
Konstante	1.490	0.022	66.34	0.000
				n=432; r^2_{adj}=0.94

Tabelle A11: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Biologie

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	0.213	0.032	6.69	0.000
Mathematik Lehramt	0.681	0.032	21.44	0.000
Chemie	0.224	0.032	7.05	0.000
Psychologie	0.076	0.032	2.40	0.017
Maschinenbau	0.458	0.032	14.42	0.000
VWL	1.061	0.032	33.39	0.000
BWL	1.310	0.032	41.24	0.000
Soziologie Magister	0.526	0.032	16.56	0.000
Germanistik Magister	0.473	0.032	14.88	0.000
Deutsch Lehramt	0.744	0.032	23.42	0.000
Jura	1.917	0.032	60.32	0.000
Konstante	1.414	0.022	62.94	0.000
				n=432; r^2_{adj}=0.94

Tabelle A12: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Chemie

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-0.011	0.032	-0.36	0.719
Mathematik Lehramt	0.457	0.032	14.39	0.000
Biologie	-0.224	0.032	-7.05	0.000
Psychologie	-0.148	0.032	-4.65	0.000
Maschinenbau	0.234	0.032	7.37	0.000
VWL	0.837	0.032	26.33	0.000
BWL	1.086	0.032	34.19	0.000
Soziologie Magister	0.302	0.032	9.51	0.000
Germanistik Magister	0.249	0.032	7.83	0.000
Deutsch Lehramt	0.520	0.032	16.37	0.000
Jura	1.693	0.032	53.27	0.000
Konstante	1.638	0.022	72.91	0.000
				n=432; r^2_{adj}=0.94

Tabelle A13: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Mathematik Lehramt

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik	-0.469	0.032	-14.75	0.000
Chemie	-0.457	0.032	-14.39	0.000
Biologie	-0.681	0.032	-21.44	0.000
Psychologie	-0.605	0.032	-19.04	0.000
Maschinenbau	-0.223	0.032	-7.02	0.000
VWL	0.380	0.032	11.95	0.000
BWL	0.629	0.032	19.80	0.000
Soziologie Magister	-0.155	0.032	-4.88	0.000
Germanistik Magister	-0.208	0.032	-6.56	0.000
Deutsch Lehramt	0.063	0.032	1.98	0.049
Jura	1.236	0.032	38.88	0.000
Konstante	2.095	0.022	93.26	0.000
				n=432; r^2_{adj}=0.94

Tabelle A14: OLS-Regression der durchschnittlichen Abschlussnote auf Studiengangdummys - Referenzkategorie Mathematik Diplom

AV: Durchschnittliche Abschlussnote	Koeffizient	Standardfehler	t-Statistik	P>|t|
Mathematik Lehramt	0.469	0.032	14.75	0.000
Chemie	0.011	0.032	0.36	0.719
Biologie	-0.213	0.032	-6.69	0.000
Psychologie	-0.136	0.032	-4.29	0.000
Maschinenbau	0.246	0.032	7.73	0.000
VWL	0.848	0.032	26.70	0.000
BWL	1.098	0.032	34.55	0.000
Soziologie Magister	0.314	0.032	9.87	0.000
Germanistik Magister	0.260	0.032	8.19	0.000
Deutsch Lehramt	0.531	0.032	16.73	0.000
Jura	1.704	0.032	53.63	0.000
Konstante	1.627	0.022	72.40	0.000
				n=432; r^2_{adj}=0.94

Tabelle A15: ANOVA der individuellen Abschlussnote mit dem Faktor Studiengang für 1967-1997

Jahr	F-Wert	Signifikanz
1967	72.14	0.000
1968	76.06	0.000
1969	87.61	0.000
1970	88.51	0.000
1971	110.26	0.000
1972	111.95	0.000
1973	112.79	0.000
1974	123.48	0.000
1975	142.04	0.000
1976	127.33	0.000
1977	174.07	0.000
1978	167.82	0.000
1979	166.72	0.000
1980	208.24	0.000
1981	156.78	0.000
1982	208.27	0.000
1983	181.21	0.000
1984	169.35	0.000
1985	167.52	0.000
1986	216.74	0.000
1987	232.49	0.000
1988	280.73	0.000
1989	282.75	0.000
1990	318.01	0.000
1991	341.12	0.000
1992	360.46	0.000
1993	426.71	0.000
1994	408.68	0.000
1995	436.00	0.000
1996	333.44	0.000
1997	260.22	0.000

Tabelle A16: Kreuztabelle Klassierungen Anzahl signifikant differenter Jahre vs. Dauer der signifikanten Differenz

Gesamt / Dauer	1	2	3	4	5
1	DLA-MLA DLA-SozMA MLA-SozMA SozMA-GerMA SozMA-MaB GerMA-MaB Mat-Che Psy-Bio	MLA-GerMA SozMA-Mat SozMA-Mat SozMA-Che SozMA-Che Mat-Psy	GerMA-Mat	VWL-MLA MLA-Mat MLA-Che SozMA-Psy SozMA-Bio MaB-Che	DLA-Mat VWL-DLA
2		MLA-MaB	GerMA-Mat Mat-Bio Mat-Bio	MaB-Mat MaB-Mat	VWL-SozMA VWL-DLA
3			Che-Psy	VWL-MLA MLA-Che MLA-Mat SozMA-Psy	VWL-SozMA VWL-DLA
4			GerMA-Che*	DLA-GerMA DLA-MaB MLA-Psy MLA-Bio SozMA-Bio MaB-Che Che-Bio	DLA-Mat GerMA-Bio
5				BWL-VWL* GerMA-Psy*	Jur-ALLE BWL-ALLE außer VWL VWL-GerMA VWL-MaB VWL-Mat VWL-Che VWL-Psy VWL-Bio DLA-Che DLA-Psy DLA-Bio MaB-Psy MaB-Bio

* Die höhere Einstufung in der Klasse der Dauer als in der Klasse der Gesamtzahl kommt durch den Ausnahmefall eines höheren Wertes für die Periodendauer als für die Gesamtzahl zustande. Dieser ergibt sich durch die Berücksichtigung einzelner nicht signifikanter Jahre zwischen zwei signifikanten Jahren in der Berechnung der Periodendauern.

Tabelle A17: Einteilung der Paarvergleiche nach Verhältnis Anzahl vs. Dauer signifikant differenter Jahre in drei Cluster

Cluster 1 (hoch/hoch) n=35	Cluster 2 n=32	Cluster 3 (niedrig/niedrig) n=14
Jura-BWL	VWL-DeutschLA	DeutschLA-MathematikLA
Jura-VWL	VWL-DeutschLA	DeutschLA-SoziologieMA
Jura-DeutschLA	VWL-DeutschLA	MathematikLA-SoziologieMA
Jura-MathematikLA	VWL-MatheLA	MathematikLA-GermanistikMA
Jura-SoziologieMA	VWL-MatheLA	SoziologieMA-GermanistikMA
Jura-GermanistikMA	VWL-SoziologieMA	SoziologieMA-Maschinenbau
Jura-Maschinenbau	VWL-SoziologieMA	SoziologieMA-Mathematik
Jura-Mathematik	DeutschLA-GermanistikMA	SoziologieMA-Mathematik
Jura-Chemie	DeutschLA-Maschinenbau	SoziologieMA-Chemie
Jura-Psychologie	DeutschLA-Mathematik	SoziologieMA-Chemie
Jura-Biologie	MathematikLA-Maschinenbau	GermanistikMA-Maschinenbau
BWL-VWL	MathematikLA-Mathematik	Mathematik-Chemie
BWL-DeutschLA	MathematikLA-Mathematik	Mathematik-Psychologie
BWL-MathematikLA	MathematikLA-Chemie	Psychologie-Biologie
BWL-SoziologieMA	MathematikLA-Chemie	
BWL-GermanistikMA	MathematikLA-Psychologie	
BWL-Maschinenbau	MathematikLA-Biologie	
BWL-Mathematik	SoziologieMA-Psychologie	
BWL-Chemie	SoziologieMA-Psychologie	
BWL-Psychologie	SoziologieMA-Biologie	
BWL-Biologie	SoziologieMA-Biologie	
VWL-Germanistik	GermanistikMA-Mathematik	
VWL-Maschinenbau	GermanistikMA-Mathematik	
VWL-Mathematik	GermanistikMA-Chemie	
VWL-Chemie	Maschinenbau-Mathematik	
VWL-Psychologie	Maschinenbau-Mathematik	
VWL-Biologie	Maschinenbau-Chemie	
DeutschLA-Mathematik	Maschinenbau-Chemie	
DeutschLA-Chemie	Mathematik-Biologie	
DeutschLA-Psychologie	Mathematik-Biologie	
DeutschLA-Biologie	Chemie-Psychologie	
GermanistikMA-Psychologie	Chemie-Biologie	
GermanistikMA-Biologie		
Maschinenbau-Psychologie		
Maschinenbau-Biologie		

Tabelle A18: Vergleich der Stichprobennoten mit Hitpass/Trosien (1987) und mit Wissenschaftsrat (2003; 2007; 2012)

	BWL (HT+WR)	BWL HoNo	VWL (HT+WR)	VWL HoNo	DeutschLA (HT+WR)	DeutschLA (HoNo)	Mathema-tikLA (HT+WR)	Mathema-tikLA (Ho-No)	Chemie (HT+WR)	Chemie HoNo	Mathema-tik (HT+WR)	Mathema-tik HoNo	Psychologie (HT+WR)	Psychologie HoNo
1953	2.45	XX	2.63	2.56	2.69	xx	2.64	xx	2.38*	2.00*	XX	1.50	2.09	XX
1963	2.80*	2.99*	3.10	3.05	2.91*	2.63*	2.75*	2.25*	1.96*	1.80*	2.30*	2.17*	2.25	2.23
1973	2.54*	2.86*	2.77*	2.64*	2.45	2.35	2.7*	2.3*	1.80	1.70	1.67	1.65	1.64	1.54
1983	2.59*	2.76*	2.63	2.59	2.65*	2.38*	2.77*	2.25*	1.90*	1.79*	1.76	1.76	1.57*	1.42*
1996	2.50*	2.70*	2.40	2.47	xx	xx	xx	xx	1.60	1.63	1.60	1.54	1.40	1.50
1998	2.50*	2.62*	2.40	2.37	xx	xx	xx	xx	1.50	1.54	1.50	1.50	1.40*	1.51*
2000	2.4*	2.65*	2.4	2.37	2.0	1.99	2.2*	1.82*	1.5	1.43	1.5*	1.71*	1.4	1.47
2005	2.15*	2.47*	2.13	2.18	1.96	1.92	2.01	2.1	1.66*	1.41*	1.58*	1.39*	1.59*	1.40*
2010	2.3	2.32	2.3	2.23	2.0*	1.59*	2.1	2.11	1.7*	1.45*	1.6*	1.44*	1.6*	1.40*

*Differenz>0.10 bzw. <-0.10

Tabelle A19: Vergleich der Stichprobendaten mit FDZ Daten für alle Hochschulen im aggregierten Querschnitt (ungewichtet gemittelt über 15/16 Jahre, 1995/6-2010)

Studiengang	Stichprobe HoNo	Bundesweit (Universitäten/TH)	Differenz
Psychologie	1.44	1.47	-0,03
Germanistik Magister	1.80	1.83	-0,03
Mathematik	1.50	1.53	-0,03
Biologie	1.36	1.40	-0,04
Soziologie Magister	1.78	1.83	-0,05
VWL	2.28	2.33	-0,05
Maschinenbau	1.85	1.92	-0,07
Chemie	1.49	1.60	-0,11
BWL	2.50	2.37	+0,13
Mathematik Lehramt	1.94	2.10	-0,16
Deutsch Lehramt	1.91	2.07	-0,16

Abbildung A1: Vergleich der Stichprobennoten mit Hitpass/Trosien (1987) und mit Wissenschaftsrat (2003; 2007; 2012) im Längsschnitt - Teil1

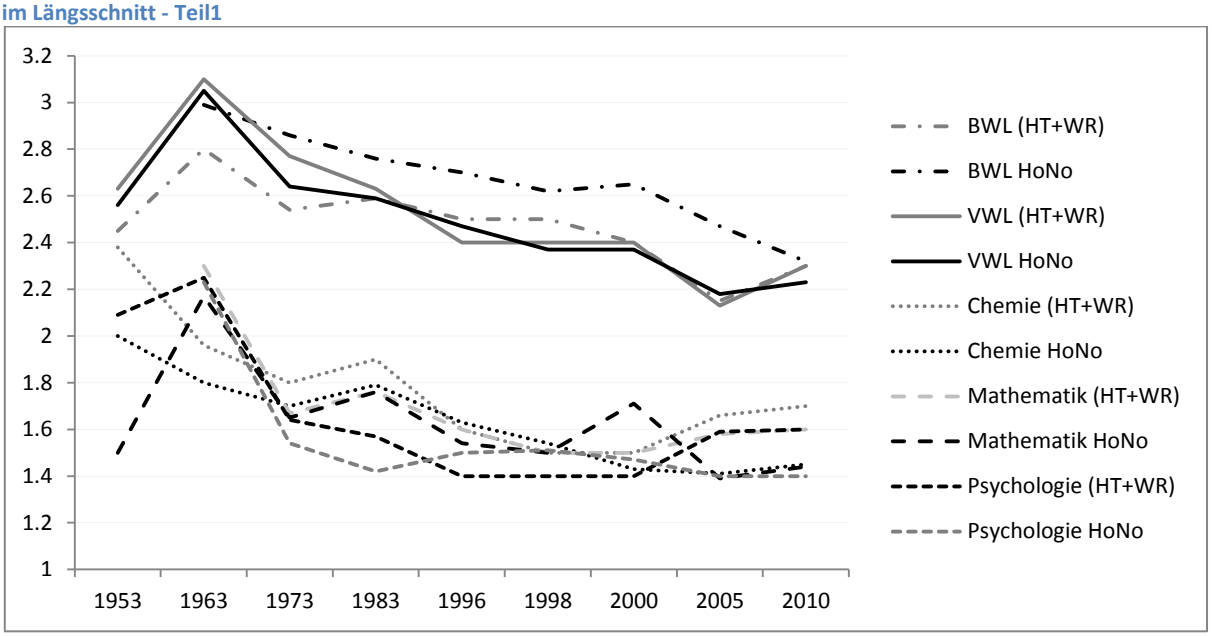


Abbildung A2: Vergleich der Stichprobennoten mit Hitpass/Trosien (1987) und mit Wissenschaftsrat (2003; 2007; 2012) im Längsschnitt – Teil2

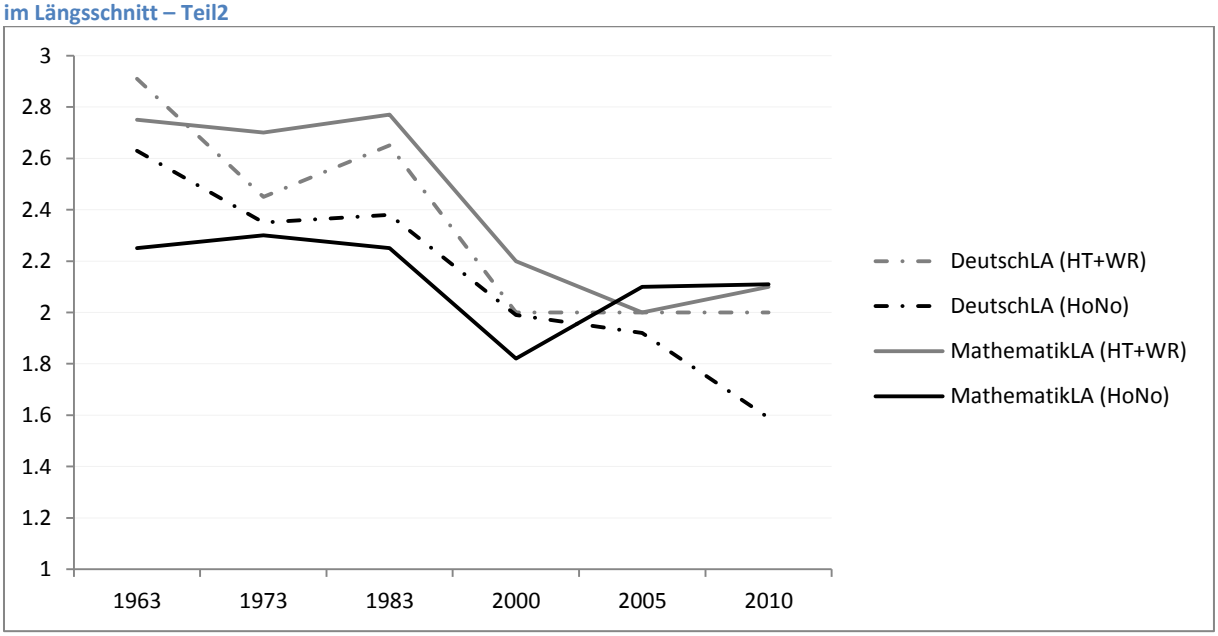


Abbildung A3: Vergleich der Stichprobennoten mit FDZ Daten für alle Hochschulen bundesweit im Längsschnitt (LOWESS 0.3) - Teil1

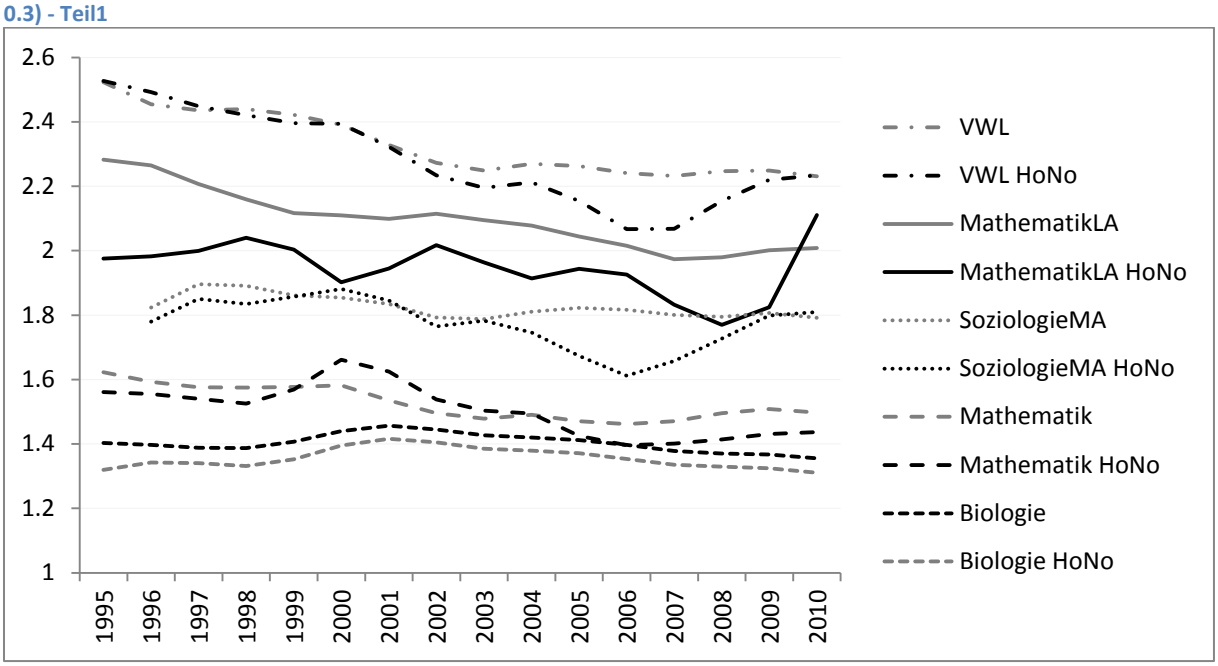


Abbildung A4: Vergleich der Stichprobennoten mit FDZ Daten für alle Hochschulen bundesweit im Längsschnitt (LOWESS 0.3) - Teil2

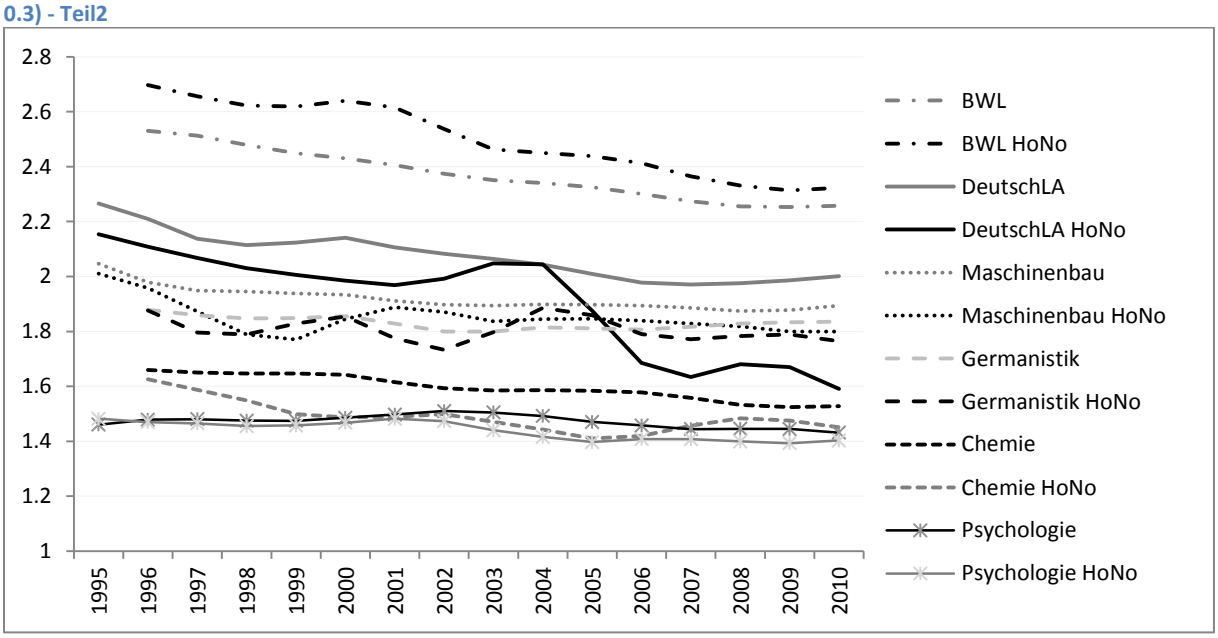


Tabelle A20: Effektstärke der Verbesserung in den einzelnen Verbesserungsphasen

Studiengang	Effektstärke (Verbesserungsphasen)
Biologie Diplom	0.922 (1967-1973)
Psychologie Diplom	1.285 (1965-1971) 0.325 (1979-1982)
VWL Diplom	0.594 (1967-1973) 0.477 (1982-1990) 0.542 (2001-2006)
Mathematik Diplom	0.848 (1963-1971) 0.524 (1985-2002)
Deutsch Lehramt	0.503 (1965-1970) 1.034 (1986-2006)
Mathematik Lehramt	0.437 (1965-1972) 0.820 (1989-2009)
BWL Diplom	0.357 (1965-1971) 0.870 (1984-2009)
Chemie Diplom	0.471 (1958-1971) 0.378 (1987-2006)

Abbildung A5: Spektraldichte der zyklischen Zeitreihenkomponente – Mathematik Diplom

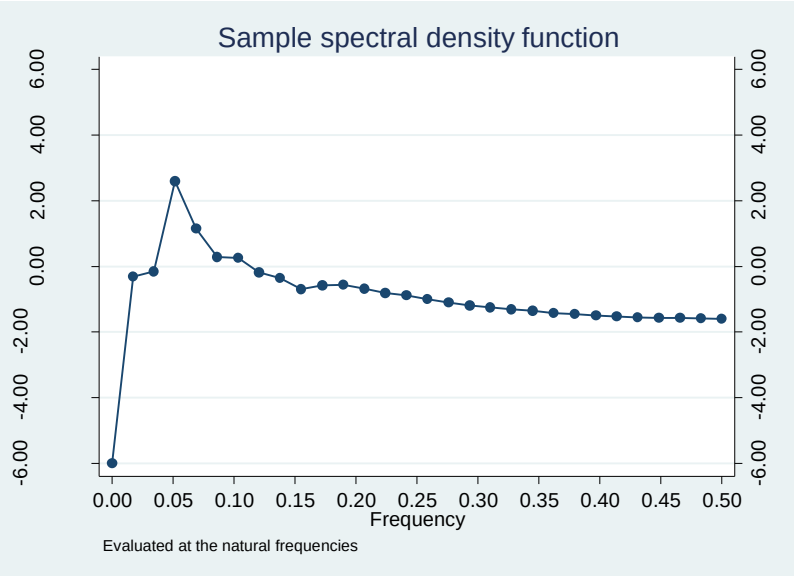


Abbildung A6: Spektraldichte der zyklischen Zeitreihenkomponente - Mathematik Lehramt

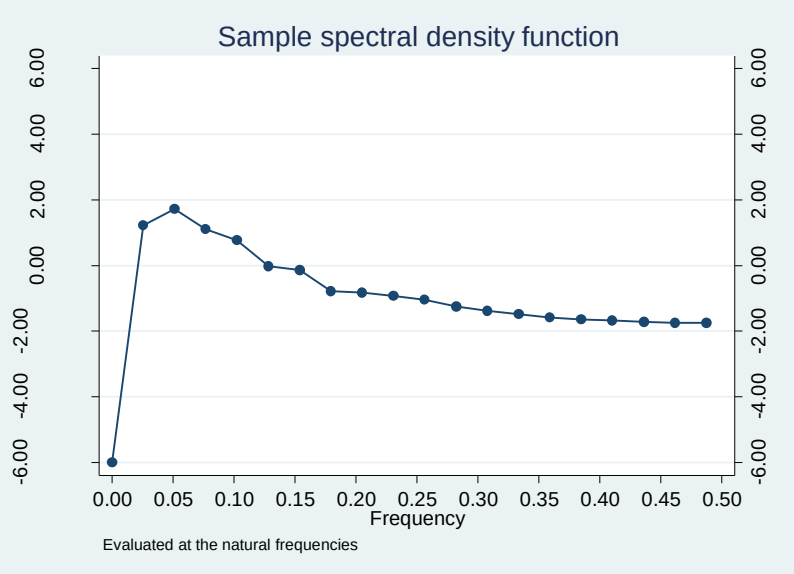


Abbildung A7: Spektraldichte der zyklischen Zeitreihenkomponente - Chemie

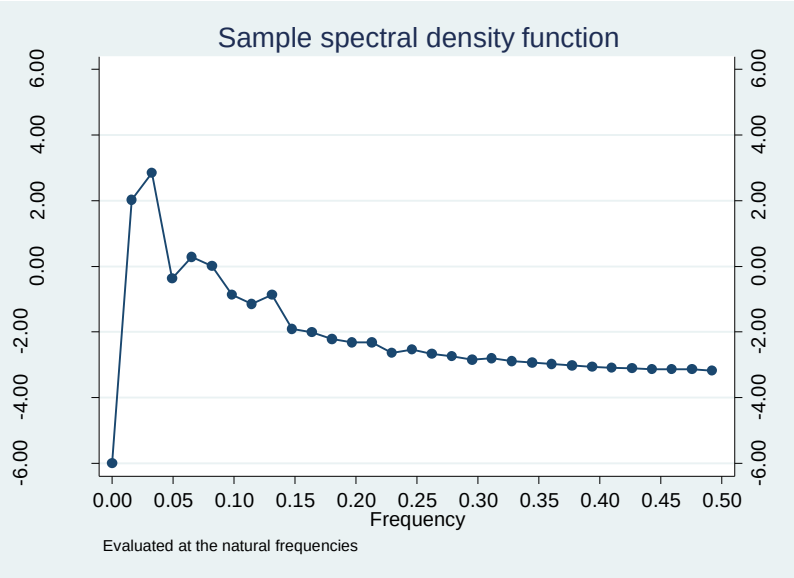


Abbildung A8: Spektraldichte der zyklischen Zeitreihenkomponente - Biologie

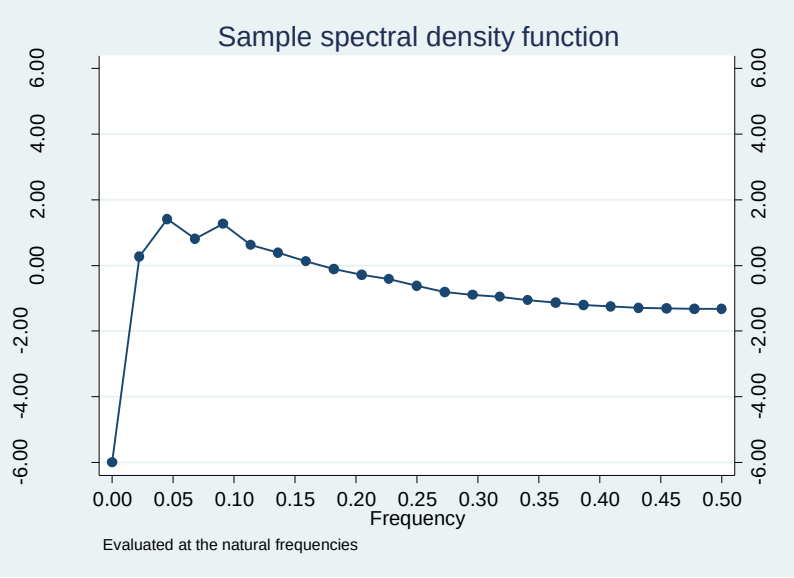


Abbildung A9: Spektraldichte der zyklischen Zeitreihenkomponente - VWL

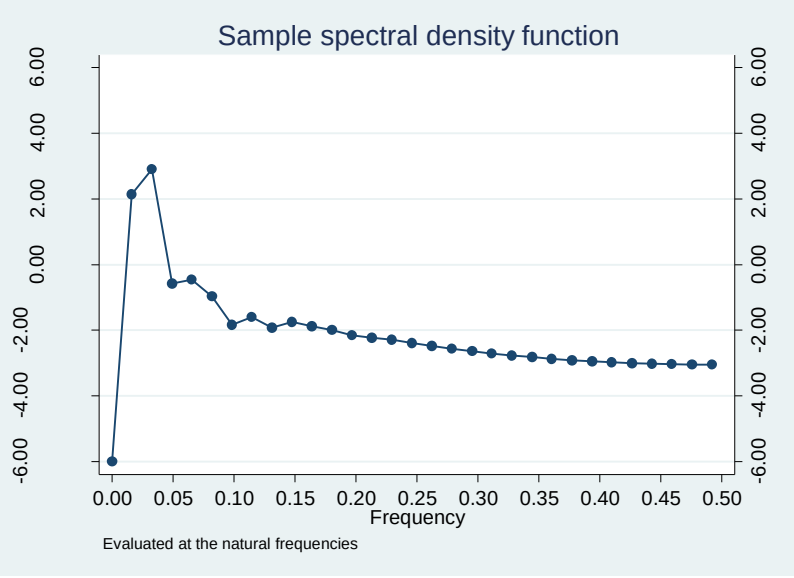


Abbildung A10: Spektraldichte der zyklischen Zeitreihenkomponente - BWL

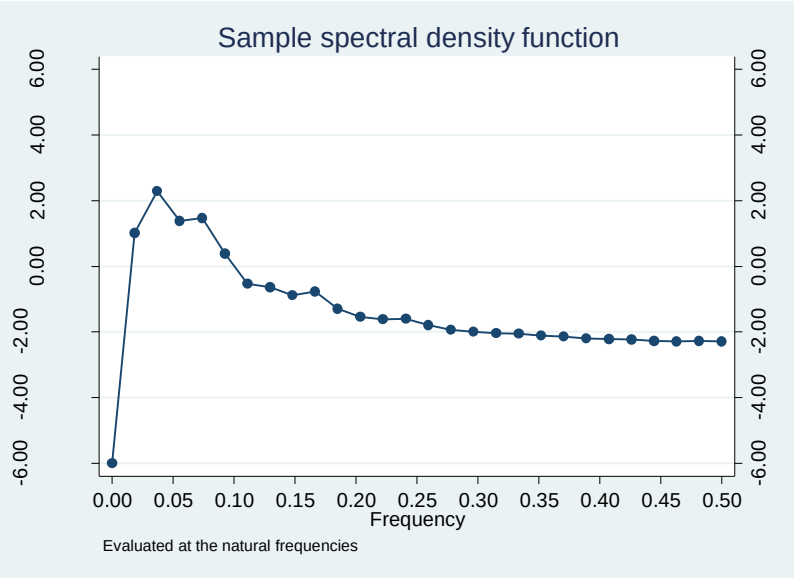


Abbildung A11: Spektraldichte der zyklischen Zeitreihenkomponente - Psychologie

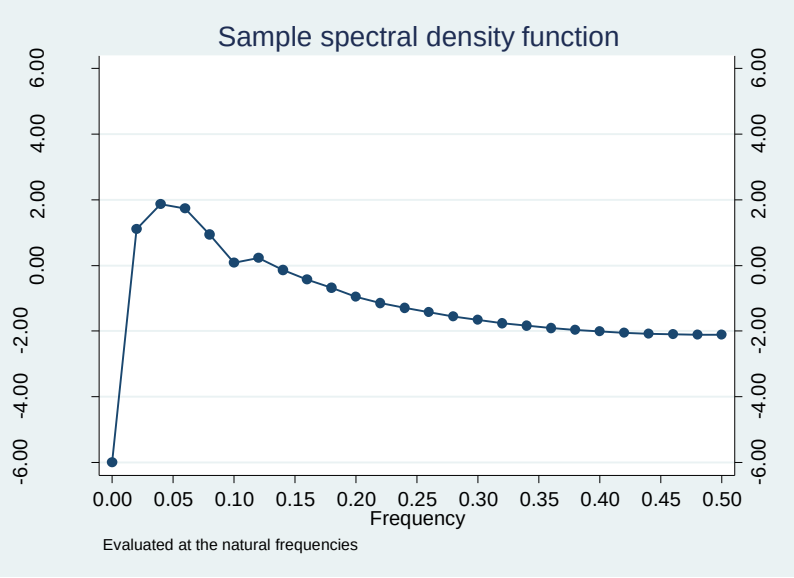


Abbildung A12: Spektraldichte der zyklischen Zeitreihenkomponente - Deutsch Lehramt

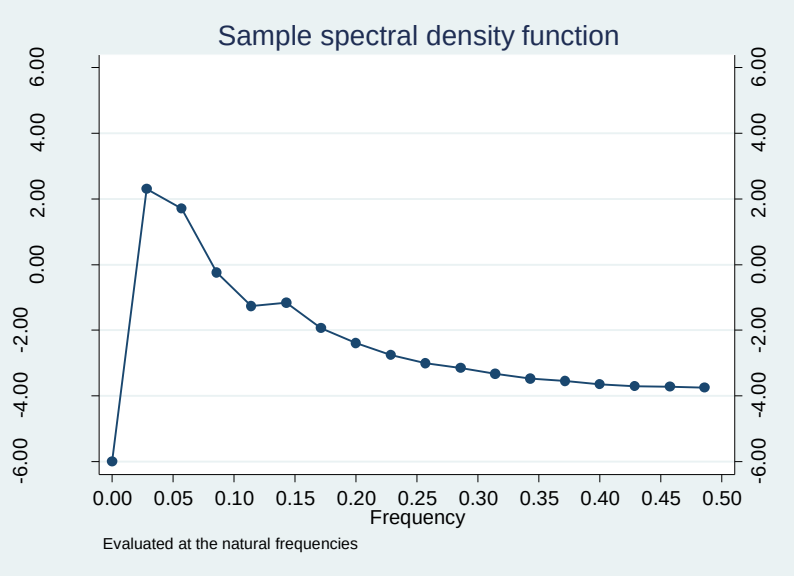


Abbildung A13: Durchschnittliche Abschlussnoten an den Hochschulen in Mathematik Diplom im Zeitverlauf

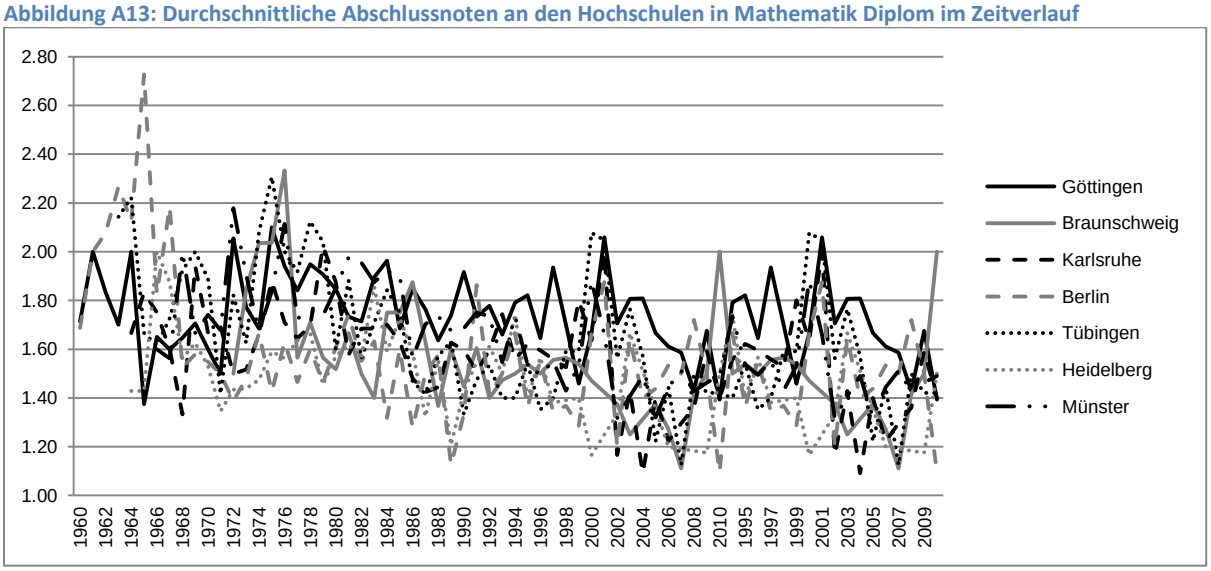


Abbildung A14: Durchschnittliche Abschlussnoten an den Hochschulen in Mathematik Lehramt im Zeitverlauf

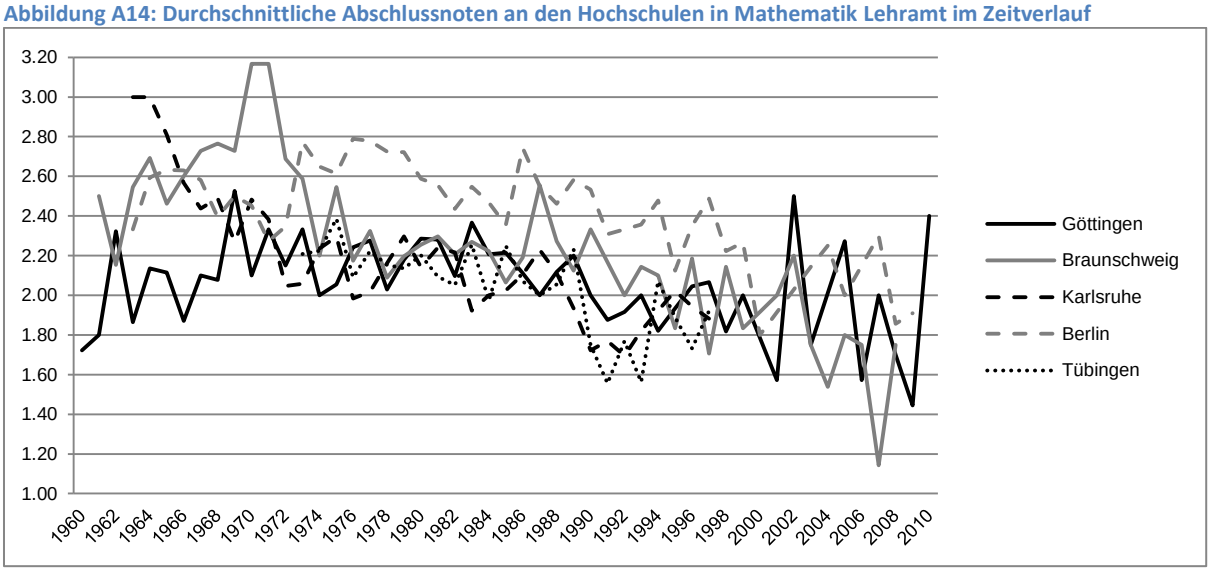


Abbildung A15: Durchschnittliche Abschlussnoten an den Hochschulen in Chemie Diplom im Zeitverlauf

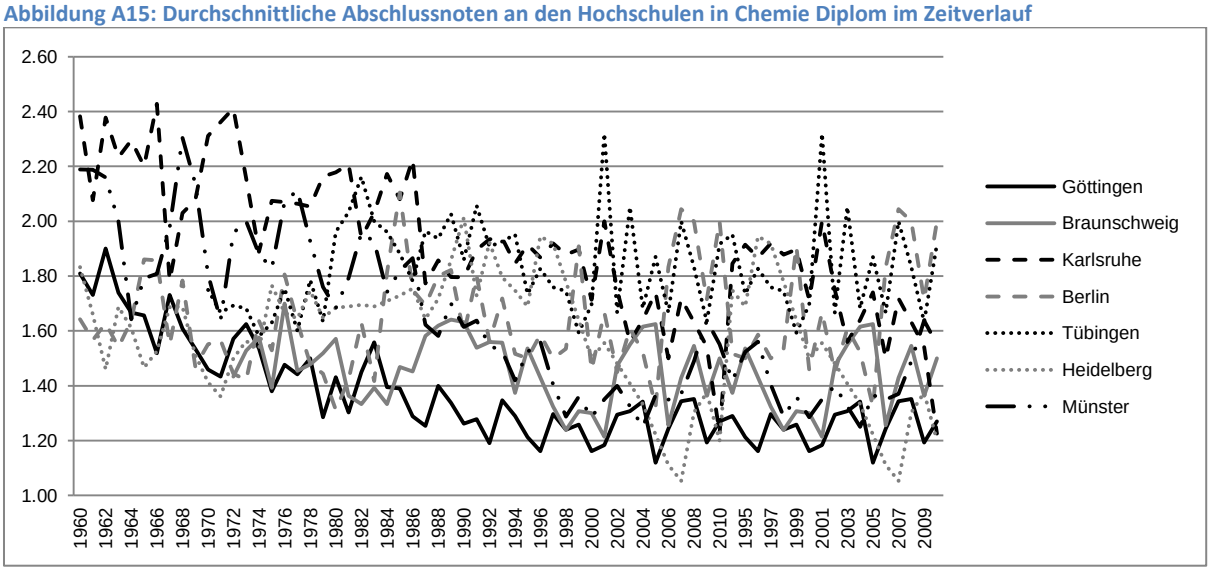


Abbildung A16: Durchschnittliche Abschlussnoten an den Hochschulen in Biologie Diplom im Zeitverlauf

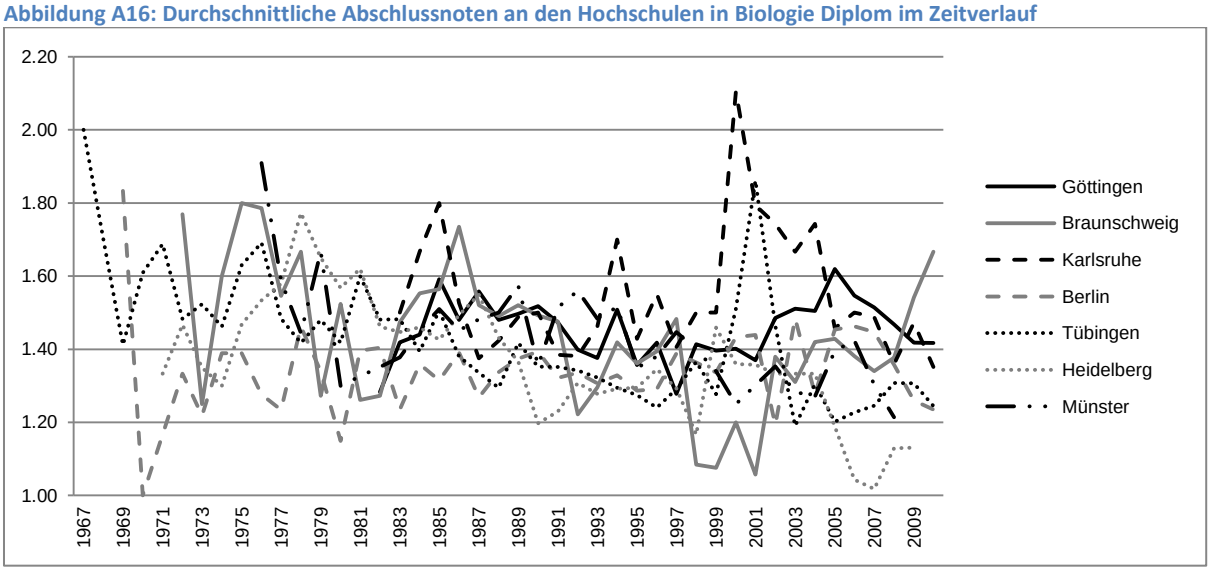


Abbildung A17: Durchschnittliche Abschlussnoten an den Hochschulen in Psychologie Diplom im Zeitverlauf

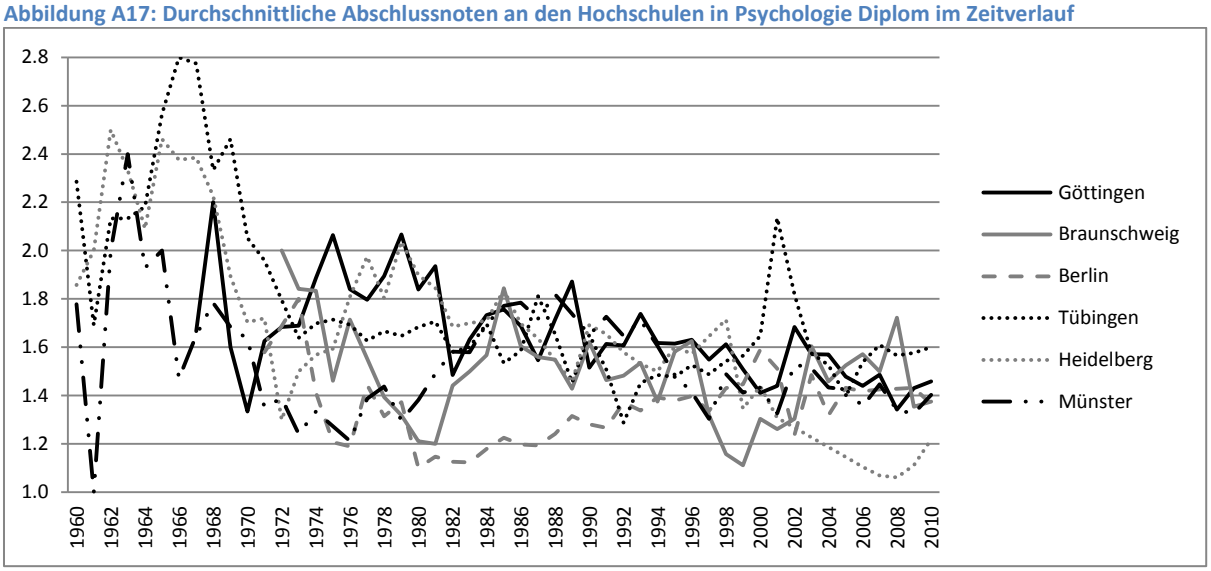


Abbildung A18: Durchschnittliche Abschlussnoten an den Hochschulen in VWL Diplom im Zeitverlauf

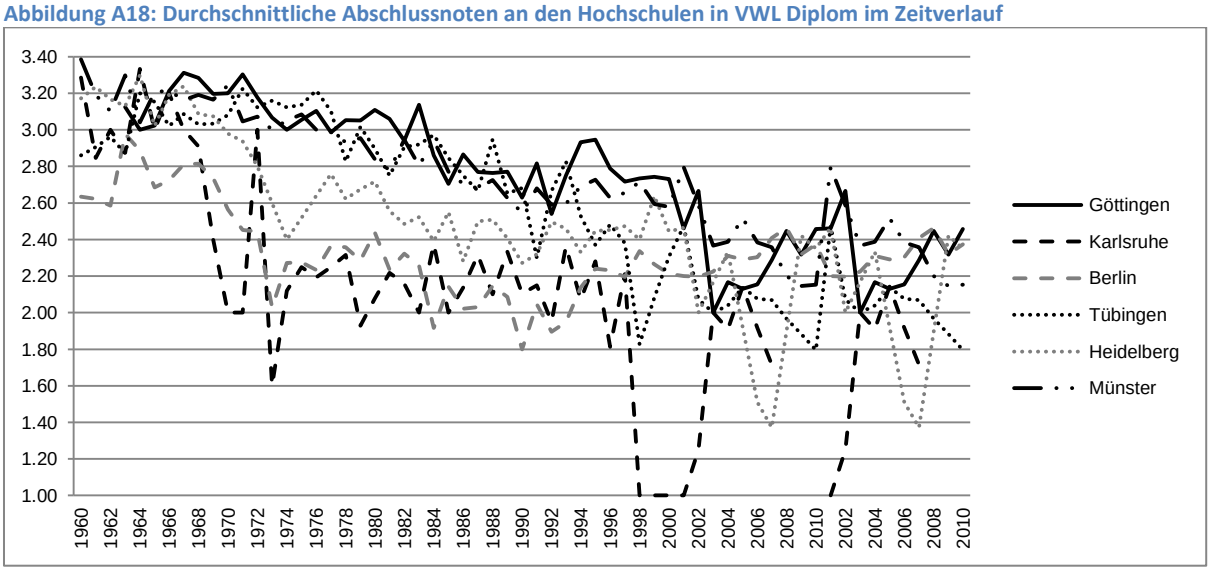


Abbildung A19: Durchschnittliche Abschlussnoten an den Hochschulen in BWL Diplom im Zeitverlauf

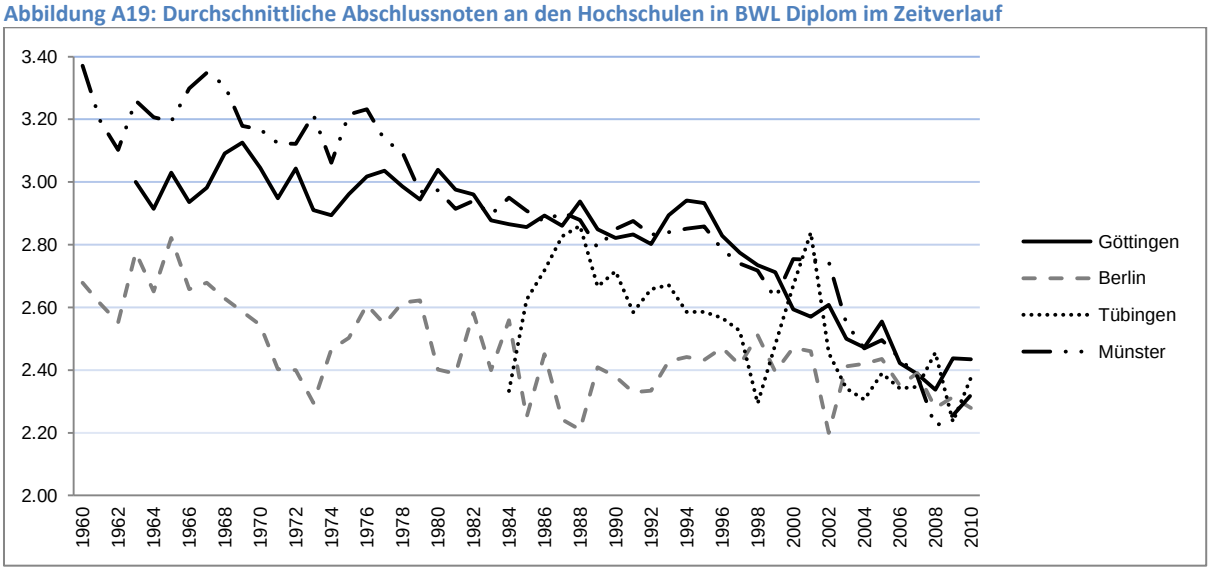


Abbildung A20: Durchschnittliche Abschlussnoten an den Hochschulen in Soziologie Magister im Zeitverlauf

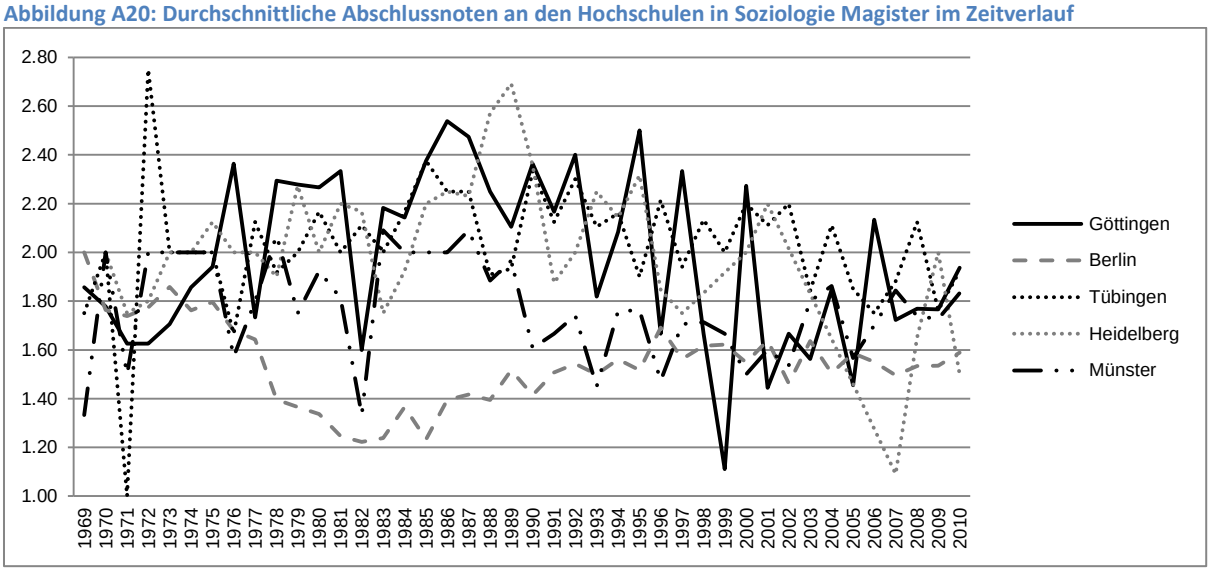


Abbildung A21: Durchschnittliche Abschlussnoten an den Hochschulen in Germanistik Magister im Zeitverlauf

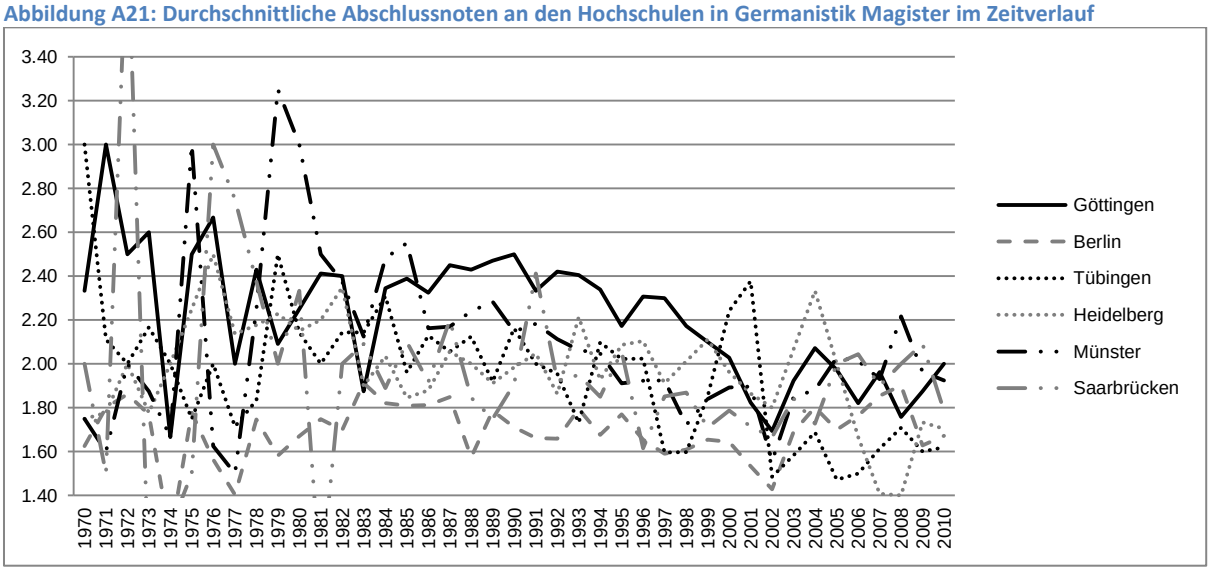


Abbildung A22: Durchschnittliche Abschlussnoten an den Hochschulen in Deutsch Lehramt im Zeitverlauf

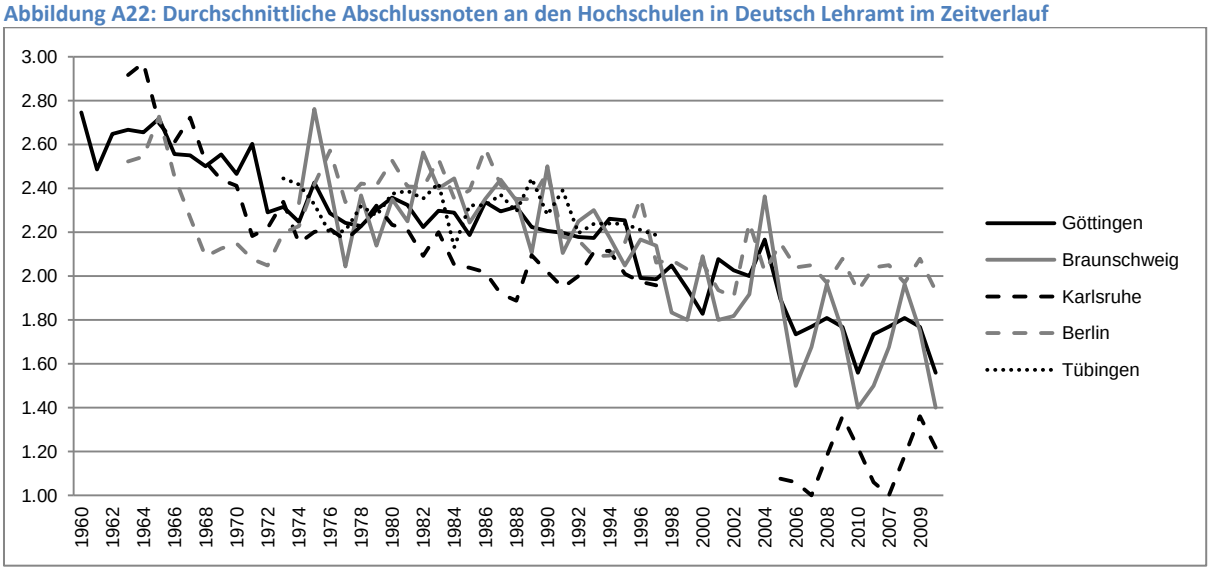
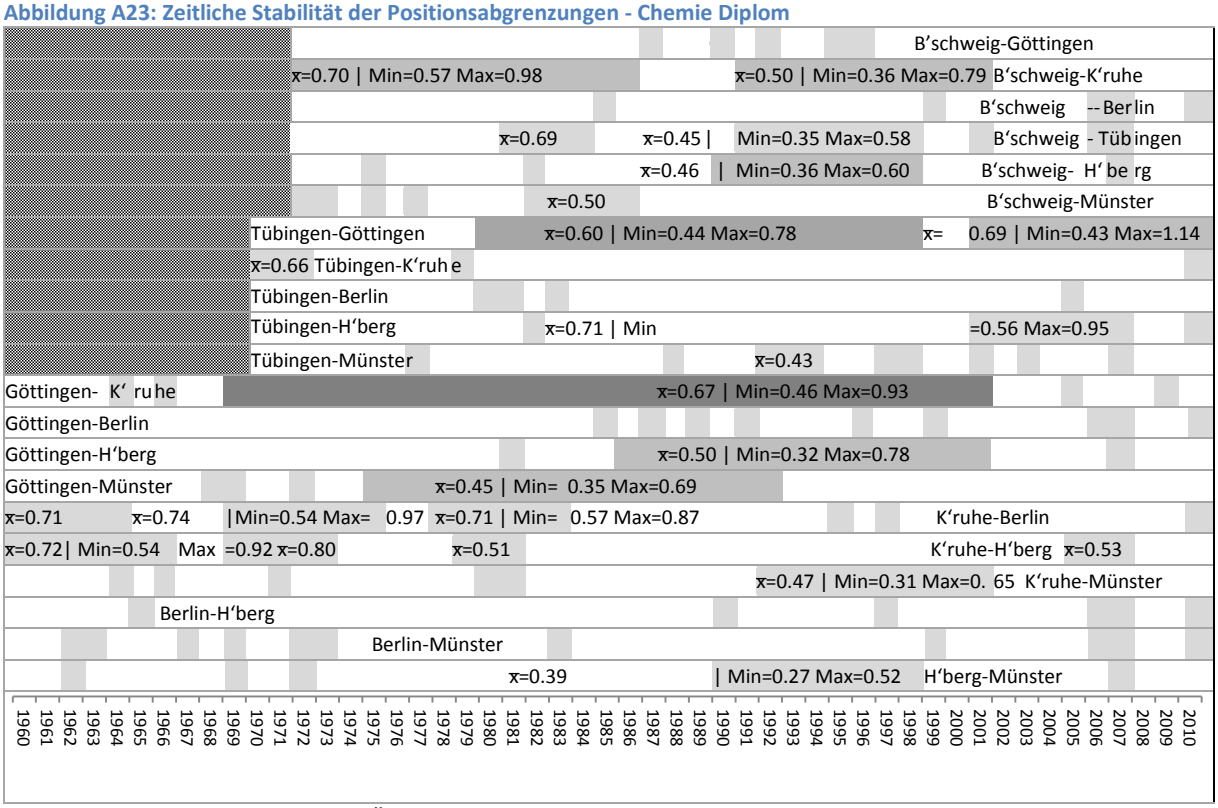


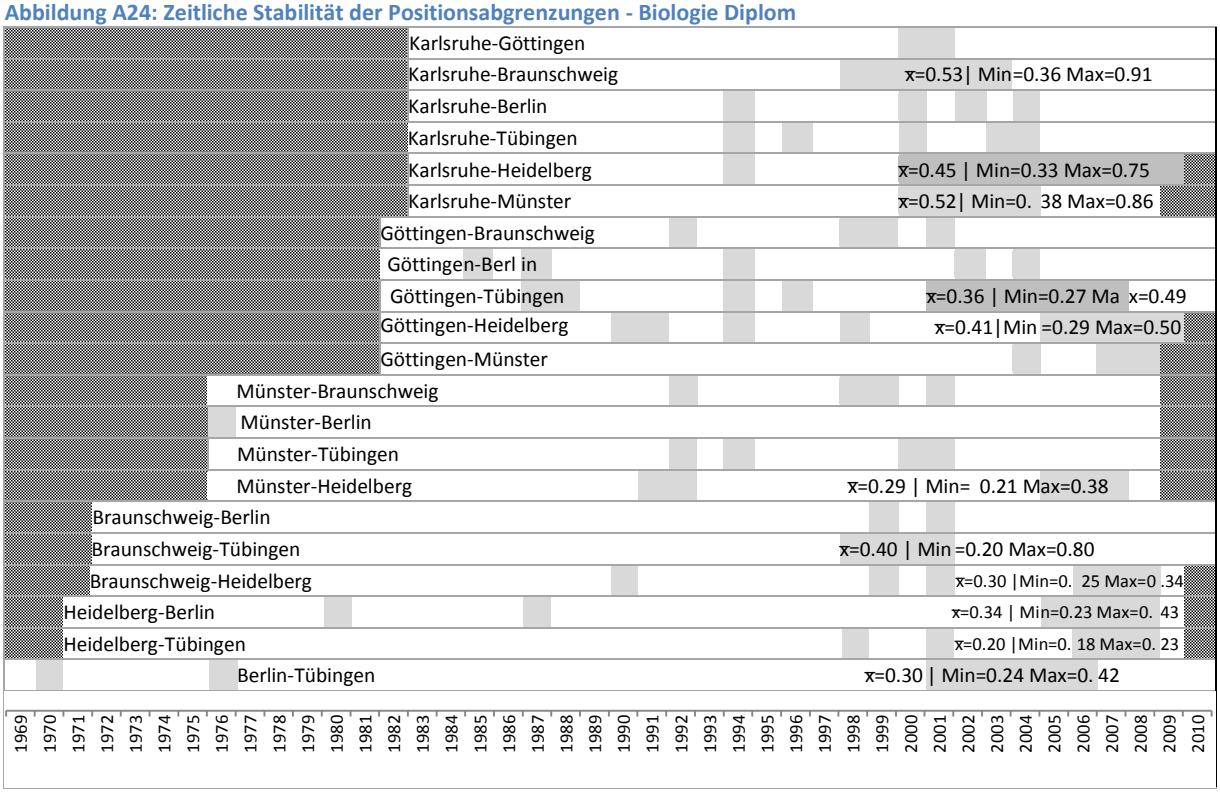
Abbildung A23: Zeitliche Stabilität der Positionsabgrenzungen - Chemie Diplom



Minimal- und Maximalwerte sind der Übersichtlichkeit halber nur bei Perioden > 5 Jahre Dauer eingetragen

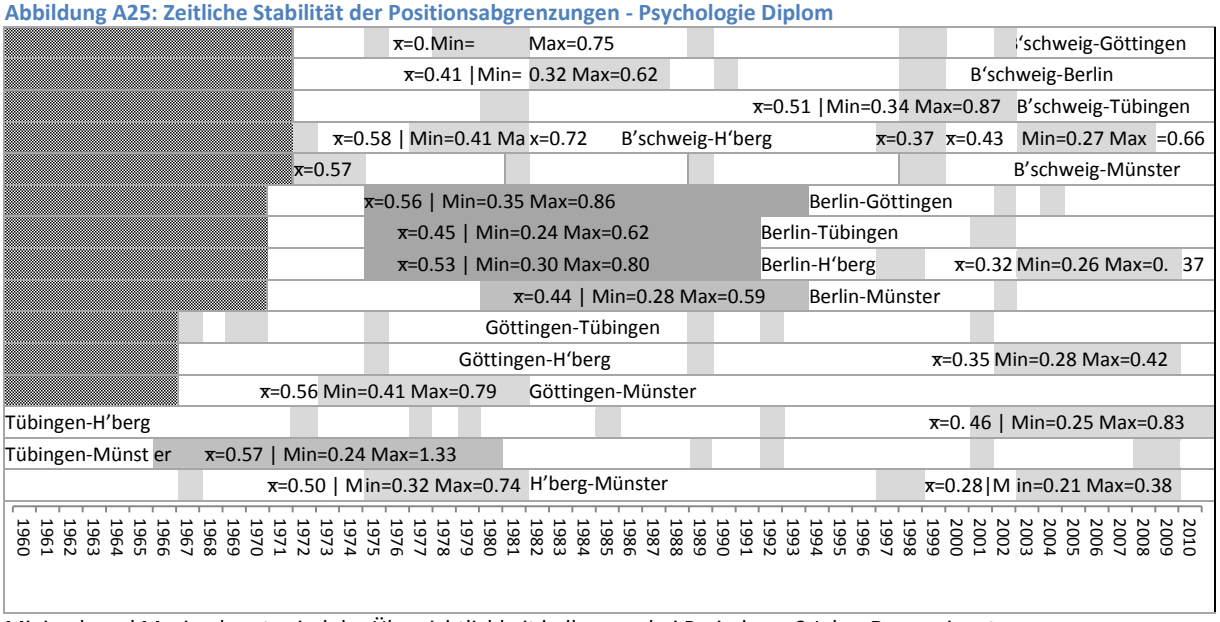
kein stabiler Unterschied	kurzfristig stabiler Unterschied	mittelfristig stabiler Unterschied	längerfristig stabiler Unterschied	langfristig stabiler Unterschied	Keine Daten vorhanden

Abbildung A24: Zeitliche Stabilität der Positionsabgrenzungen - Biologie Diplom



kein stabiler Unterschied	kurzfristig stabiler Unterschied	mittelfristig stabiler Unterschied	längerfristig stabiler Unterschied	langfristig stabiler Unterschied	Keine Daten vorhanden

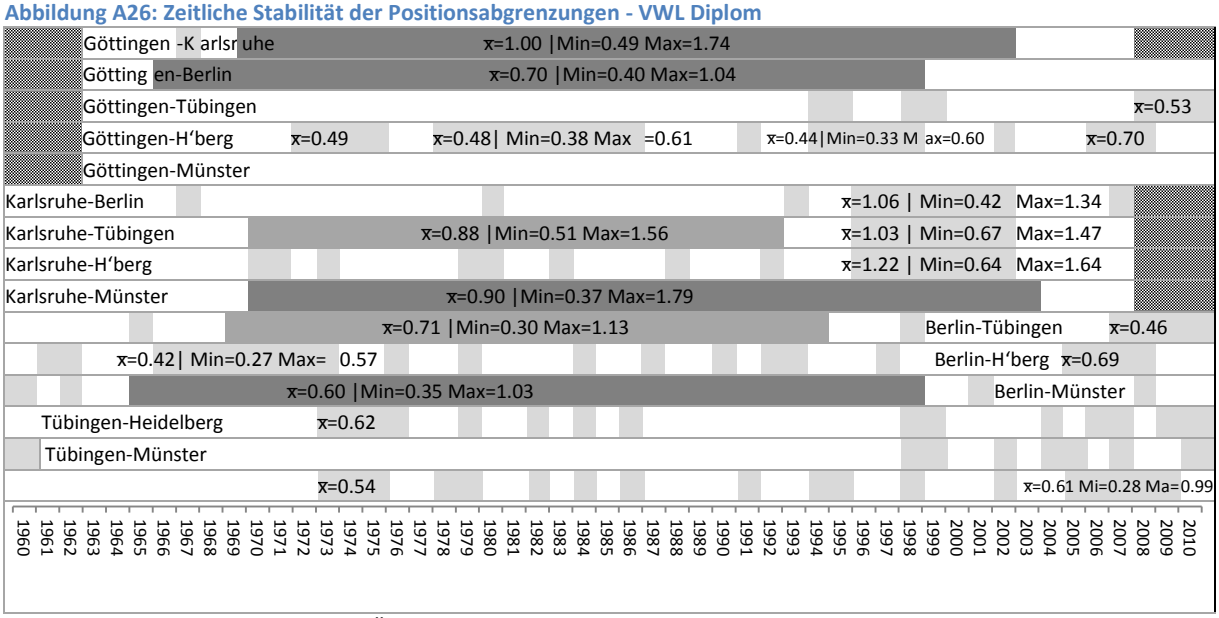
Abbildung A25: Zeitliche Stabilität der Positionsabgrenzungen - Psychologie Diplom



Minimal- und Maximalwerte sind der Übersichtlichkeit halber nur bei Perioden > 3 Jahre Dauer eingetragen

kein stabiler Unterschied	kurzfristig stabiler Unterschied	mittelfristig stabiler Unterschied	längerfristig stabiler Unterschied	langfristig stabiler Unterschied	Keine Daten vorhanden

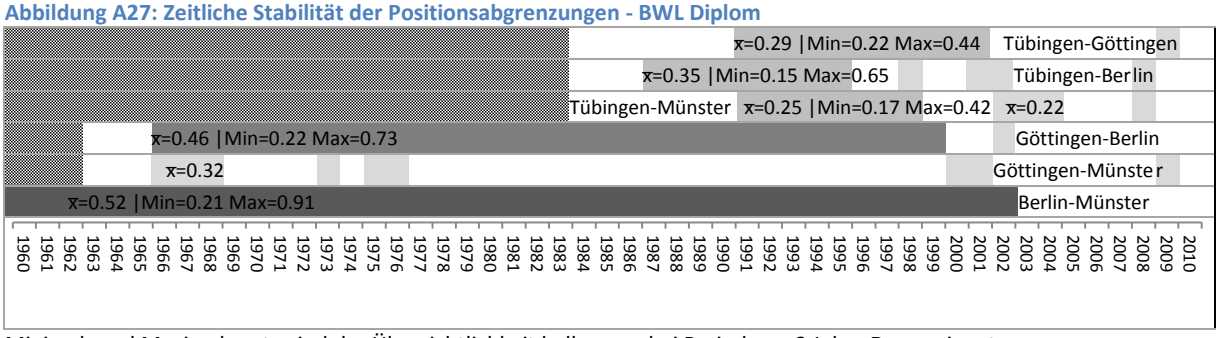
Abbildung A26: Zeitliche Stabilität der Positionsabgrenzungen - VWL Diplom



Minimal- und Maximalwerte sind der Übersichtlichkeit halber nur bei Perioden > 4 Jahre Dauer eingetragen

kein stabiler Unterschied	kurzfristig stabiler Unterschied	mittelfristig stabiler Unterschied	längerfristig stabiler Unterschied	langfristig stabiler Unterschied	Keine Daten vorhanden

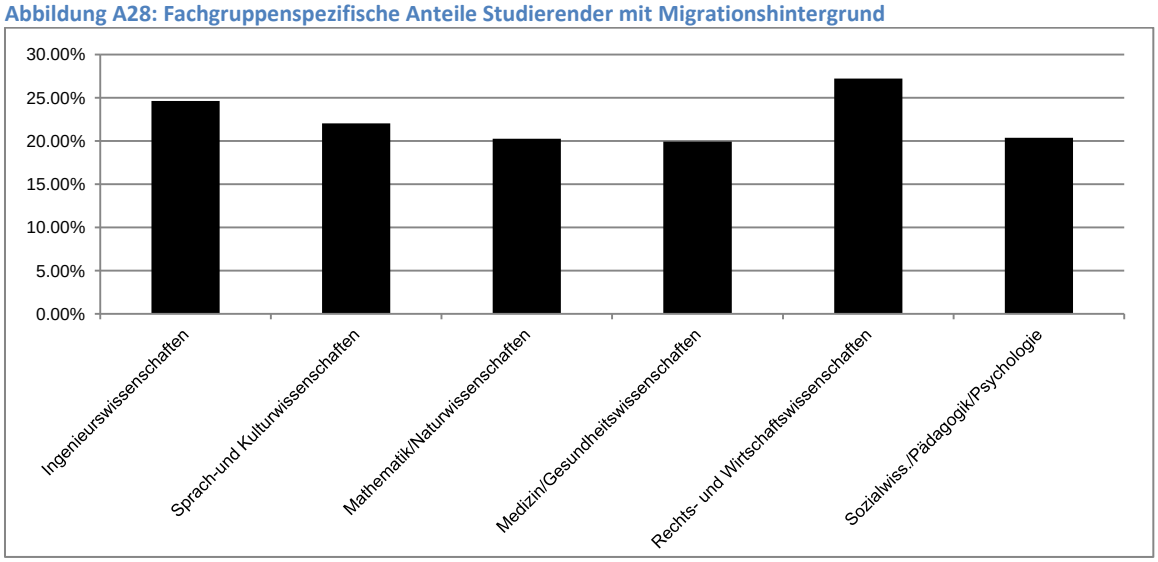
Abbildung A27: Zeitliche Stabilität der Positionsabgrenzungen - BWL Diplom



Minimal- und Maximalwerte sind der Übersichtlichkeit halber nur bei Perioden > 3 Jahre Dauer eingetragen

kein stabiler Unterschied	kurzfristig stabiler Unterschied	mittelfristig stabiler Unterschied	längerfristig stabiler Unterschied	langfristig stabiler Unterschied	Keine Daten vorhanden

Abbildung A28: Fachgruppenspezifische Anteile Studierender mit Migrationshintergrund



Quelle: Middendorff et al. (2013): Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2012, eigene Berechnungen

Abbildung A29: Partielle Autokorrelationsfunktion der trendbereinigten Abschlussnoten in BWL (Prä-Intervention)

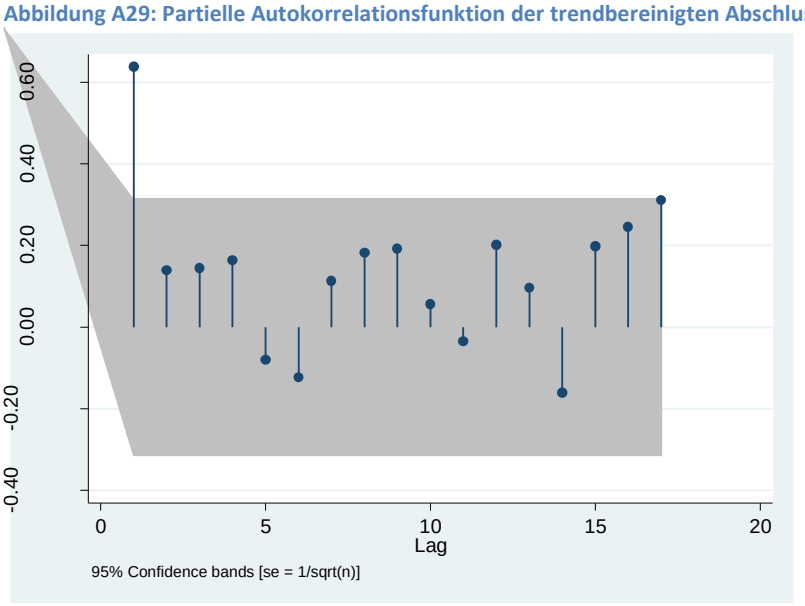


Abbildung A30: Autokorrelationsfunktion der trendbereinigten Abschlussnoten in BWL (Prä-Intervention)

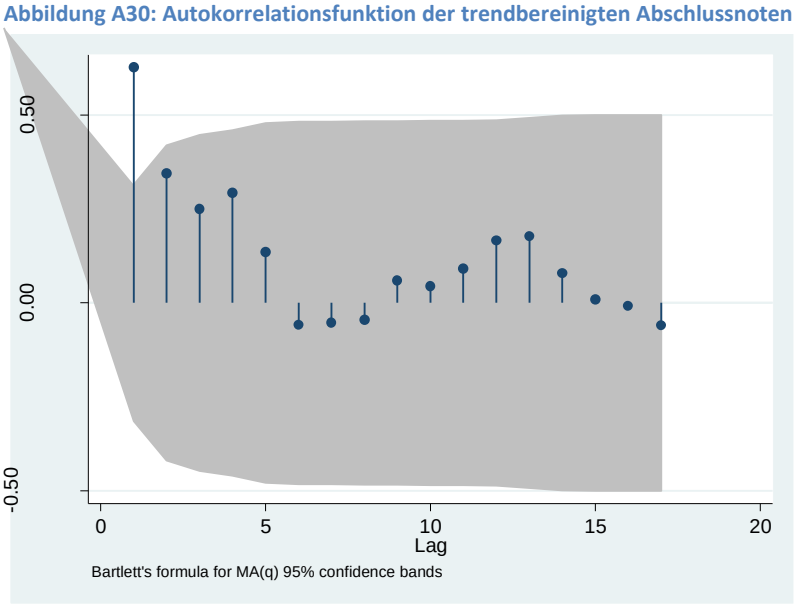


Abbildung A31: Partielle Autokorrelationsfunktion der trendbereinigten Abschlussnoten in VWL an der Universität Göttingen (Prä-Intervention)

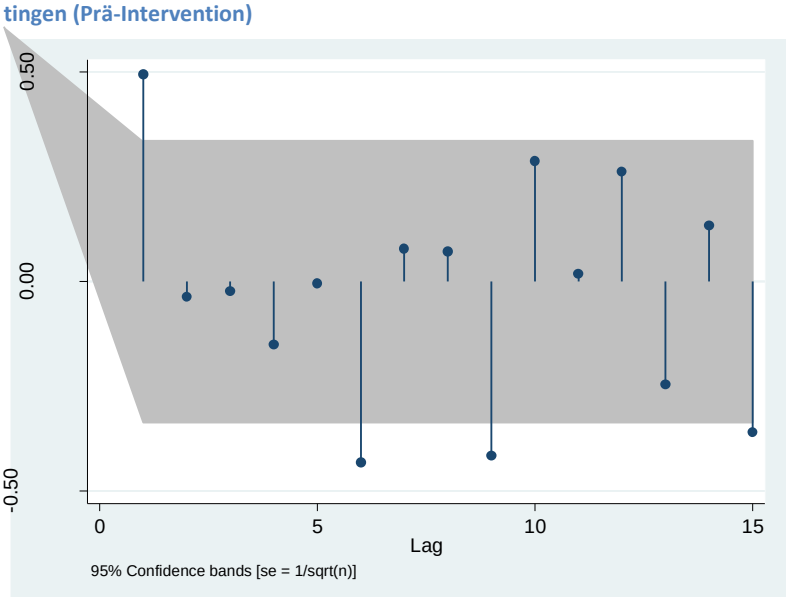


Abbildung A32: Autokorrelationsfunktion der trendbereinigten Abschlussnoten in VWL an der Universität Göttingen (Prä-Intervention)

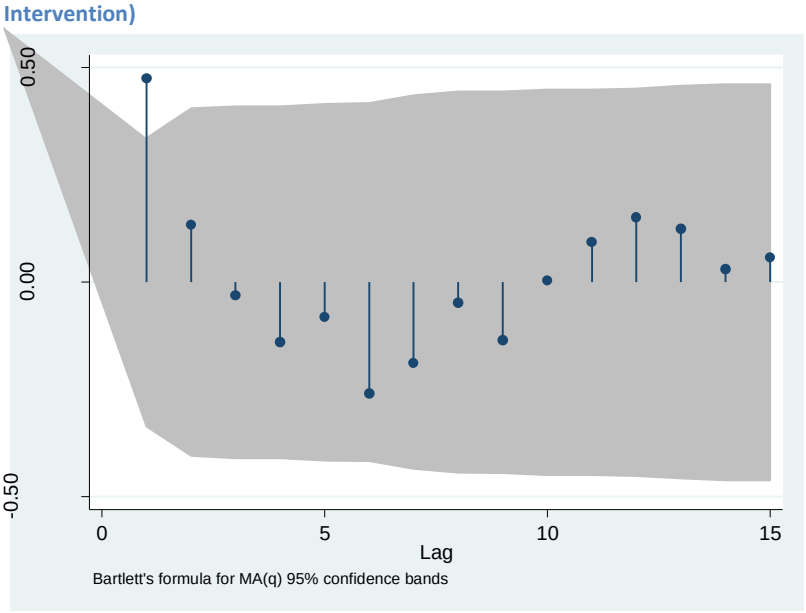


Tabelle A21: Lineare Regression der Zwischenprüfungsnote auf leistungskonforme Prüfungsbedingungen

AV: Note Zwischenprüfung	Koeffizient	Standardfehler	beta	t-Statistik	P>|t|
Note Abitur	0.385	0.011	0.358	35.37	0.000
Alter bei Studienbeginn	-0.007	0.002	-0.033	-3.33	0.001
Geschlecht weiblich[a]	-0.060	0.014	-0.043	-4.31	0.000
Index der beruflichen Position	-0.016	0.005	-0.038	-3.23	0.001
Akademiker*innenkind[b]	0.027	0.016	0.020	1.68	0.094
Stipendiat*in[c]	-0.295	0.036	-0.079	-8.13	0.000
Erhebungswelle	0.118	0.014	0.441	8.41	0.000
Erhebungswelle_quadriert	-0.011	0.001	-0.490	-9.35	0.000
Konstante	2.063	0.080		25.74	0.000
					n=9150; r^2adj=0.15

[a]Referenzkategorie: männlich [b]Referenzkategorie: Kein Akademiker*innenkind [c]Referenzkategorie: kein*e Stipendiat*in

Abbildung A33: Anzahl Prüflinge z-standardisiert, trendbereinigt, geglättet

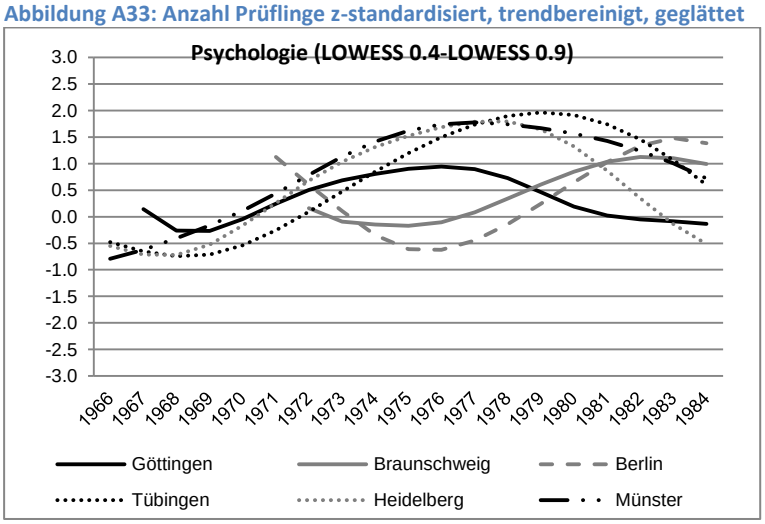


Abbildung A34: Anzahl Prüflinge z-standardisiert

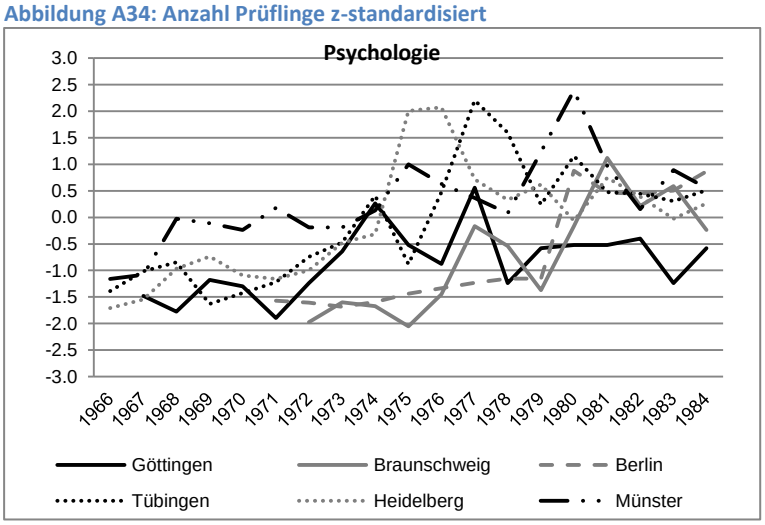


Tabelle A22: P-W-Regression der nicht geglätteten Abschlussnoten auf die Prüfungszahlen für Studiengänge mit Fachkonjunktur

	Koeffizient	Standardfehler	t-Statistik	P>|t
Note Mathematik Diplom				
Anzahl Prüflinge (Lead 3)	-0.469	0.251	-1.87	0.069
Jahr	-0.028	0.015	-1.89	0.066
Konstante	56.77	29.89	1.90	0.065
				D-W= 1.98; r^2_{adj}=0.35; n=42
Note Mathematik Lehramt				
Anzahl Prüflinge (Lead 3)	-0.161	0.093	-1.73	0.092
Jahr	-0.053	0.008	-6.98	0.000
Konstante	106.3	15.21	6.98	0.000
				D-W= 1.82; r^2_{adj}=0.54; n=43
Note Chemie Diplom				
Anzahl Prüflinge (Lead 0)	0.179	0.082	2.20	0.033
Jahr	-0.064	0.006	-10.3	0.000
Konstante	127.2	12.39	10.3	0.000
				D-W= 2.00; r^2_{adj}=0.67; n=51
Note Biologie Diplom				
Anzahl Prüflinge (Lead 0)	-0.288	0.165	-1.74	0.090
Jahr	-0.052	0.015	-3.56	0.001
Konstante	103.7	29.10	3.56	0.001
				D-W= 1.82; r^2_{adj}=0.59; n=39
Note Psychologie Diplom				
Anzahl Prüflinge (Lead 4)	-0.252	0.107	-2.35	0.025
Jahr	-0.038	0.013	-2.84	0.008
Konstante	75.63	26.66	2.84	0.008
				D-W= 2.12; r^2_{adj}=0.27; n=38
Note VWL Diplom				
Anzahl Prüflinge (Lead 4)	-0.122	0.069	-1.78	0.085
Jahr	-0.081	0.007	-11.3	0.000
Konstante	161.4	14.30	11.3	0.000
				D-W= 1.92; r^2_{adj}=0.79; n=35
Note BWL Diplom (bis 2000[a])				
Anzahl Prüflinge (Lead 0)	0.284	0.074	3.84	0.000
Jahr	-0.078	0.006	-14.1	0.000
Konstante	155.0	11.03	14.1	0.000
				D-W= 2.31; r^2_{adj}=0.82; n=48
Note Deutsch Lehramt (bis 1998[a])				
Anzahl Prüflinge (Lead 5)	-0.189	0.077	-2.44	0.021
Jahr	-0.059	0.008	-7.38	0.000
Konstante	117.8	15.92	7.40	0.000
				D-W= 1.97; r^2_{adj}=0.64; n=31

[a] Die Beschränkungen für BWL und Deutsch Lehramt wurden gewählt, weil im folgenden Zeitraum die Datengrundlage (FDZ-Daten) zu unsicher ist. Deutsch LA = Berlin, Göttingen, Karlsruhe.

Tabelle A23 85: P-W-Regression der nicht geglätteten Abschlussnoten auf die Prüfungszahlen für Studiengänge ohne Fachkonjunktur

	Koeffizient	Standardfehler	t-Statistik	P>|t
Note Germanistik Magister Göttingen				
Anzahl Prüflinge (Lead 0)	0.462	0.136	3.40	0.002
Jahr	-0.078	0.011	-7.10	0.000
Konstante	155.7	21.92	7.10	0.000
				D-W= 1.89; r^2_{adj}=0.57; n=42
Note Germanistik Magister Berlin (OLS Regression)				
Anzahl Prüflinge (Lead 0)	0.309	0.177	1.75	0.088
Jahr	-0.007	0.015	-0.46	0.647
Konstante	13.55	29.32	0.46	0.647
				D-W= 1.80; r^2_{adj}=0.03; n=41
Note Germanistik Magister Heidelberg				
Anzahl Prüflinge (Lead 0)	-0.351	0.195	-1.80	0.080
Konstante	-0.051	0.233	-0.22	0.828
				D-W= 1.88; r^2_{adj}=0.08; n=41
Note Germanistik Magister Saarbrücken				
Anzahl Prüflinge (Lead 1)	-0.235	0.088	-2.67	0.011
Jahr	0.002	0.007	0.30	0.768
Konstante	-4.124	13.95	-0.30	0.769
				D-W= 1.16; r^2_{adj}=0.28; n=40
Note Soziologie Magister Berlin				
Anzahl Prüflinge (Lead 0)	-0.231	0.107	-2.15	0.037
Jahr	-0.039	0.017	-2.37	0.022
Konstante	78.31	33.03	2.37	0.022
				D-W= 2.65; r^2_{adj}=0.16; n=49
Note Soziologie Magister Heidelberg				
Anzahl Prüflinge (Lead 1)	-0.524	0.154	-3.39	0.002
Jahr	0.004	0.020	0.21	0.833
Konstante	-8.628	40.78	-0.21	0.834
				D-W= 1.93; r^2_{adj}=0.22; n=40